IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 9, 2019, accepted August 4, 2019, date of publication August 8, 2019, date of current version August 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2933987

Hand Gesture Recognition Based on Active Ultrasonic Sensing of Smartphone: A Survey

ZHENGJIE WANG, YUSHAN HOU, KANGKANG JIANG, WENWEN DOU, CHENGMING ZHANG, ZEHUA HUANG, AND YINJING GUO

College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding authors: Zhengjie Wang (cieewangzj@163.com) and Yinjing Guo (gyjlwh@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61471224, in part by the Qingdao Postdoctoral Applied Research Project under Grant 2015180, and in part by the Shandong Province Key Research and Development Plan (Public Welfare Special) Project under Grant 2018GHY115022.

ABSTRACT With the rapid development of Internet of Things, hand gesture recognition has drawn wide attention in the field of ubiquitous computing because it provides us with simple and natural human-computer interaction mode. Among these various implementations, hand gesture recognition using ultrasonic signals of smartphone has become a hot research topic due to its various advantages. In this paper, we consider the smartphone as an active sonar sensing system to identify hand movements. Specifically, the speakers emit ultrasonic signal and the microphone on the same phone receives the changed echo affected by hand movements. This paper investigates the state-of-the-art hand gesture applications and presents a comprehensive survey on the characteristics of studies using the active sonar sensing system. Firstly, we review the existing research of hand gesture recognition based on acoustic signals. After that, we introduce the characteristics of ultrasonic signal and describe the fundamental principle of hand gesture recognition. Then, we focus on the typical methods used in these studies and present a detailed analysis on signal generation, feature extraction, preprocessing, and recognition methods. Next, we investigate the state-of-the-art ultrasonic-based applications of hand gesture recognition using smartphone and analyze them in detail from dynamic gesture recognition and hand tracking. Afterwards, we make a discussion about these systems from signal acquisition, signal processing, and performance evaluation to obtain some insight into development of the ultrasonic hand gesture recognition system. Finally, we conclude by discussing the challenges, insight, and open issues involved in hand gesture recognition based on ultrasonic signal of the smartphone.

INDEX TERMS Doppler effect, hand gesture recognition, smartphone, ultrasonic signal.

I. INTRODUCTION

With the rapid development of Internet of Things (IoT), the demand for pervasive sensing is continually increasing because it can provide us with more information and facilitate the development of ubiquitous applications. These applications usually measure and amass the sensing data by using various signals, such as video [1], sound [2], radio frequency (RF) [3], and light [4]. These signals have distinct characteristics and can be leveraged according to the requirements of applications. We can categorize these applications into two groups: environment sensing and target sensing. The aim of the former is to collect and assess environmental information such as temperature of environment or weather information, etc. and the purpose of the latter is to measure and monitor the target state, such as crowdsensing, elderly health monitoring, human activity recognition, etc. In this paper, we concentrate on the sound signal and its applications for sensing active target. The acoustic sensing techniques have been extensively studied and applied in various scenarios because sound signal has many advantages, such as simple waveform, good physical features, low deployment cost. Recently, many novel features of sound signal have been explored and quite a number of the acoustic signal-based applications have been developed in various aspects. We can categorize these applications into quite a number of groups based on their purposes, such as speech recognition [5], activity recognition [6]–[13], health monitoring [14]–[29], lip reading [30], identity authentication [31]–[38], attack and defense [39]–[43], finger tapping detection [44], localization [45]–[53], indoor

The associate editor coordinating the review of this article and approving it for publication was Hong-Ning Dai.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

mapping [54], [55], acoustic imaging [56], context awareness [57]–[60], touch detection [61]–[63], speed measurement [64], multi-device interaction [65]–[68], safe driving and walking [69]–[72], fault detection [73], [74], indoor navigation [75], and motion tracking [76]–[80], etc.

Among the current acoustic signal-based recognition applications, human hand gesture recognition has become a hot research topic and attracted more attention because it can provide many wonderful applications such as game control, device unlocking, identity authentication, and information input, etc. The basic idea of these approaches is that hand movement may disturb the sound signal propagation and lead to signal changes. The receiver can capture the changed signal caused by hand movement and the system can identify the changes by comparing the echo with the original signal. Then the signal changes could be analyzed to recognize hand motions. In recent years, quite a number of studies of hand gesture recognition have increasingly emerged. Based on their purposes, these applications usually employ different hardware equipment to emit and receive sound signal, including smartwatch [81]–[83], computer [84]–[90], specially designed device [91]–[96], and transceiver in environment [97], [98].

Among these applications, some systems require participants to wear sensors and some systems depend on the specific sensors, which can be inconvenient under many scenarios and lead to extra deployment cost. Meanwhile, some applications use the sound devices of laptops or desktops, which brings some difficulties when deploying in some small space. Besides, some systems apply smartwatches to recognize hand postures. These methods have some disadvantages, such as extra deployment cost, inconvenience of wearing a device, requirements for customized devices, etc., which limits the application area of sound sensing.

Fortunately, with the rapid development of smartphone, its capabilities of computation and environment sensing have been strengthened increasingly. Moreover, due to the powerful function of smartphone, it has become an indispensable electronic device. As a result, we can utilize their built-in sensors, such as speakers and microphones, to develop many interesting applications using sound signals of the smartphone. The systems using smartphone as signal sensing sensor have the natural advantages, such as zero deployment cost, ubiquitously available devices, and powerful computing capability. Furthermore, these systems benefit from the good development pattern because we can implement sensing functions by deploying application programs on the smartphone. Currently, a host of studies and applications for hand recognition based on acoustic signal of smartphone have emerged. We classify these applications into two types: passive sensing system and active sensing system. To be specific, we can record the ambient sound signal using the microphone of the smartphone and recognize hand gesture using the changed signal. We call these applications as passive sensing systems because the smartphone solely captures the sound signal. Besides, the smartphone can serve as an active sonar system to emit and receive sound signal, which enables us to explore the active sonar by modulating the signal and generating appropriate waveform. We call these applications as active sensing systems because the smartphone first emits sound signal and then records the signal variation. Both two types of applications greatly extend the application range of sound signal due to the popularity of smartphones.

A. PASSIVE ACOUSTIC SENSING

The aim of passive sensing is to capture the ambient sound signals or the sound signal transmitted by other devices. Therefore, the microphones embedded in the smartphone serve as a sound input device in this scenario [78], [79], [99]–[110]. For instance, UbiK [100] leverages the dual-microphone embedded in the smartphone to collect signals and extracts the multipath fading as features. It can recognize the text inputting from the keyboard outline printed on the conventional surfaces (e.g., wood table). SoundWrite [102] and SoundWrite II [103] recognize the stroke according to the sound signal generated by moving the finger on the surface (e.g., table, paper). They leverage the microphone to capture the acoustic signal, extract frequency and time features, and recognize stroking by pattern classification. SoundWave II adds two threshold values and exploits the Mel frequency cepstral coefficient (MFCC) to extract stable features and improve noise tolerance. UbiWriter [104] utilizes the microphones built in smartphone to record the audio signal generated by handwriting and analyzes the audio signal to realize the text input.

B. ACTIVE ACOUSTIC SENSING

The meaning of active sensing is that the sound signal is emitted and received by the same smartphone. The speaker built in smartphone transmits ultrasonic signal at a given frequency and the microphone embedded in the same smartphone captures the changed ultrasonic signal affected by the hand movement around the mobile. The signal changes are leveraged to recognize hand gesture based on the pattern rules or geometric models [111]–[132]. From this view, the smartphone can be treated as an active sonar. For example, AudioGest [112] is a fine-grained hand motion recognizing system based on smartphone. This system can not only classify hand postures but also identify the speed of hand movement, the range of hand motion, and the duration of hand in the air. Specifically, AudioGest expands the number of hand gestures by taking these three factors into consideration. Different from AudioGest, R. Nandakumar *et al*. propose a finger tracking system called FingerIO [111]. This system can track the finger by using the microphone and speaker embedded in smartphone. It utilizes the orthogonal frequency division multiplexing (OFDM) technology to obtain the transmitted sound signal. Moreover, FingerIO enables finger tracking available even the mobile in a pocket.

Currently, although many acoustic-based human gesture recognition applications have been developed and applied

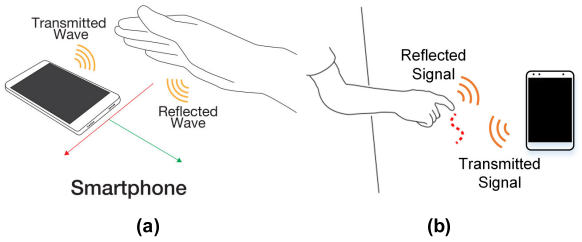


FIGURE 1. The common hand gesture scenarios based on active acoustic sensing. (a) Dynamic gesture recognition [116]; (b) Hand trajectory tracking.

in various scenarios using commodity hardware, the comprehensive survey on acoustic sensing is still very deficient. Reference [2] is the latest review on acoustic sensing based on commodity devices. However, it focuses on the application layer, processing layer, and physical layer. There is no survey on hand gesture recognition using the ultrasonic signal from smartphone. This paper concentrates on the system framework, processing techniques, and applications using ultrasonic signal from smartphone. Specifically, in this paper, we focus on the applications of hand gesture recognition which leverage the speakers built in smartphone to transmit ultrasonic signals and utilize the microphones embedded on the same smartphone to record the echoes, including dynamic hand gesture recognition and hand trajectory tracking, as shown in Fig. 1. The former focuses on the specific gestures and labels the unknown movement according to collected data. The latter tracks the movement trajectory of the hand, such as drawing the shape or alphabet. Therefore, these two applications have distinct characteristics and we will analyze them in application part. We investigate these applications and analyze their features to facilitate the development of natural and novel human-computer interface (HCI) methods.

The contributions of this paper can be summarized as follows. Firstly, we present the review of recent progress in hand gesture recognition based on ultrasonic signal of smartphone. To our best of knowledge, this paper is the first survey on the hand gesture recognition by an ultrasonic signal based on smartphones. Secondly, we give the typical framework to recognize hand gesture using the built-in speakers and microphones of smartphone. We illustrate the signal processing procedure from signal measurement to behavior recognition. It comprises ultrasonic signal collection, preprocessing, and hand gesture identification. We describe each step in detail and analyze the corresponding algorithms. Finally, we investigate the existing applications and classify them into two types according to the purpose of applications: dynamic gesture recognition and hand trajectory tracking. We make a comprehensive comparison of these applications and evaluate their performance from several aspects, e.g., experiment devices, extracted features, signal preprocessing, classification approaches.

The rest of this paper is organized as follows: we first introduce the fundamental principle and the basic system framework of the hand gesture recognition using ultrasonic signal of smartphone in Section II. Then, we analyze some essential techniques about hand gesture recognition using ultrasonic signal in Section III, including signal generation, signal analysis, signal preprocessing, and the recognition methods. After that, we present some existing applications of gesture recognition based on ultrasonic signal and evaluate their system performance from several aspects, such as experimental scenarios, conducted actions, findings, and dimensions, etc. in Section IV. We make a comprehensive discussion about these applications from signal acquisition, signal processing, and performance evaluation in Section V. At last, we present the limitations, challenges, and future directions in Section VI. Section VII is our conclusion.

II. OVERVIEW OF HAND GESTURE RECOGNITION

In this section, we present an overview of hand gesture recognition based on ultrasonic signal of smartphone, including the fundamental principle and the basic system framework. The fundamental principle describes how to the system recognize hand gesture using the changed ultrasonic signal affected by hand movement and the system framework depicts the main components of the system and interprets their functions.

A. FUNDAMENTAL PRINCIPLE

As shown in Fig. 1, it illustrates the common hand gesture recognition scenarios using smartphone. The speakers of a smartphone emit the ultrasonic signal and the microphones from the same smartphone receive the changed signal affected by human hand movement. When the scenario is empty, the received signal is similar to the emitted signal except for some effects from ambient noise. If the hand keeps stationary at a specific distance from the phone in the scenario, the received signal is blocked or reflected by hand. We can calculate the distance between the hand and the phone by leveraging the signal attenuation or reflection rules. When the hand moves, waves, draws a shape, or writes a letter, hand movements will change the signal propagation and reflection path continuously. If we want to recognize hand gesture, we can exploit the pattern classification techniques. If we want to track hand movement, we usually explore the mathematical and geometric models because we need to continuously determine the hand position with low error. Otherwise, the features of letter or shape may be destroyed, leading to the failure of recognition.

These recognition systems adopt an ultrasonic signal from smartphone because the inaudible sound can provide us with many advantages. Because of the popularity of the smartphone, we can easily utilize it as a signal measurement device, which helps us to develop widespread and daily applications without any extra deployment cost. Besides, the ultrasonic signal has been studied thoroughly and has been successfully applied in various environments. For gesture recognition applications, these systems adopt the sound signal with a frequency more than 16 kHz as the emitted signal. Since the sound signal with a frequency more than 16 kHz is generally beyond the average human's audibility [133], we can call these signals as the ultrasonic signal. And the ultrasound

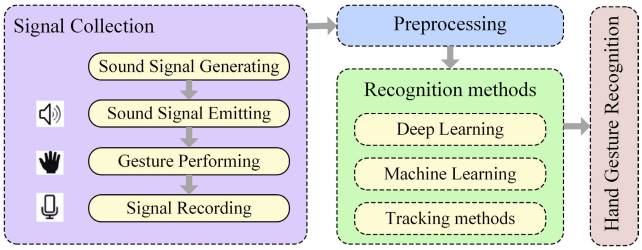


FIGURE 2. The system framework of hand gesture recognition based on active ultrasonic sensing of smartphone.

signal has drawn more attention in hand posture recognition applications due to the following reasons. Firstly, ultrasound signal has been studied thoroughly and has been widely applied to localization and other applications due to its good ranging accuracy and low-cost deployment. Secondly, ultrasound is beyond humans' hearing. Therefore, it can be used to monitor participants without disturbing his/her normal life, which enables long-term monitoring available. Thirdly, it can work well under many scenarios without light even through the wall, which provides evident benefits compared with other signals.

The features of the received sound signal include amplitude, frequency, initial phase, and propagation time. When we utilize the sound signal to recognize hand gesture, we can use the above information or their modification forms, such as channel impulse response (CIR), phase, frequency, and time of flight (ToF). Because we consider the smartphone as an active sonar system, we can develop different signal coding patterns, such as sine wave, orthogonal frequency division multiplexing (OFDM) signal, chirp signal, and binary phase shift keying (BPSK) signal. These different types of signals effectively improve recognition accuracy because they can enhance signal synchronization, suppress noise, increase resolution and sensitivity, and improve signal-noise ratio (SNR). The introduction about signal is presented in Section III.A.

B. SYSTEM FRAMEWORK

In this section, we present a typical framework of hand gesture recognition based on active ultrasonic sensing of smartphone. As shown in Fig. 2, it comprises four parts: signal collection, signal preprocessing, recognition methods, and gesture recognition. We first interpret the procedures of hand gesture identification using inaudible signal from smartphone. Then, we illustrate the specific function of each component.

Based on the requirements of design, the system first selects a suitable sound signal, including continuous wave, OFDM signal, BPSK signal, and chirp signal. Then the speakers transmit the sound, which can be implemented by playing the recorded sound file. At the same time, participants wave his/her hand to conduct some gestures or draw a shape or letter. The microphones receive the transformed signal and store it in memory.

After collecting signals, we need to analyze the component and extract useful information to identify the hand movement. The received signal comprises useful ultrasonic signal and various noises from environment and hardware which severely affect the precision of measurement data. Therefore, we must first eliminate noises and abnormal values. The simple and effective ways of noise removal are to adopt filter techniques, including low-pass filter and band-pass filter. After that, we get clear data which is used to determine movement procedure. Next, we need to obtain the start and end time of the movement to divide the data stream into segments. These segments can be fed into a classifier to recognize gestures or be used to locate hand position to implement finger tracking.

After signal preprocessing, we obtain sound segments. For dynamic gesture recognition, we consider it as a classification problem as we evaluate algorithm performance based on predefined gestures. Therefore, we can utilize machine learning algorithm and deep learning algorithm to identify hand movement because they are general classifiers and can be used at our studies. For hand tracking, we usually consider it from two aspects: localization and trajectory tracking. The former involves a distance measurement between hand and smartphone while the latter contains continuous position calculation. The hand tracking can be implemented using model-based methods, such as ToF and geometric model. Besides, if we can effectively utilize the feature of the speed of the target, the tracking accuracy can be improved because the speed of hand motion cannot change sharply. And we usually exploit geometric model to calculate distance and track hand position using phase and ToF.

III. PROCESSING TECHNIQUES

In this section, we analyze the implementation of systems from signal processing view. Specifically, we first present the signal collection, signal preprocessing, and then illustrate the gesture identification methods. The signal collection part discusses the signal waveform, modulation parameters, and signal components. The signal preprocessing part analyzes the essential approaches for obtaining the effective data. And the last part presents the implementation algorithms of gesture recognition and hand trajectory tracking.

A. SIGNAL COLLECTION

1) SIGNAL GENERATION

The human hand gesture recognition applications based on ultrasonic signal of smartphone apply the ultrasound signal to recognize the hand postures. Speaker transmits the ultrasonic signal and microphone receives the sound signal reflected by human hand movements, then the received signals are processed with some algorithms to identify hand postures.

Because we develop and deploy the recognition systems based on active sonar pattern, we can exert more control to sound signal. The characteristics of the sound signal are important criteria for designing the types of audio signals. A good sound signal can bring about a satisfactory result in noise removal and feature extraction. As a

TABLE 1. Comparison of systems, including adopted signal, extracted signal, sensors, devices, number of devices, additional sensors, and device-free.

System	Adopted signal	Extracted signal	Sensors	Devices	Number of devices	Additional sensors	Device-free
AudioGest [112]	19 kHz sine acoustic wave	Doppler shift	one speaker, one microphone	Samsung Galaxy S4	1	No	Yes
Dolphin [113]	21 kHz continuous tone	Doppler shift	one speaker, one microphone	MI One, Samsung S3	1	No	Yes
SonicOperator [117]	21 kHz sine acoustic wave	Doppler shift	one speaker, one microphone	MI Note, Vivo X7, MI 5, HTC One	1	No	Yes
UltraGesture [118]	20 kHz	CIR	two speakers, five microphones	Samsung S5, microphone-speaker kit	1	No	Yes
AirLink [119]	18.8 kHz pilot tone	Doppler shift	speakers, microphones	Samsung Galaxy Nexus, Samsung Galaxy S3	3	No	Yes
VSkin [120]	17-23 kHz ZC sequence	Phase, Amplitude	one speaker, two microphones	Samsung S5, Huawei Mate7, Samsung S7, Samsung Note3	1	No	No
ForcePhone [121]	18-24 kHz linear chirp signal	Vibration amplitude, Signal change ratio	one speaker, one microphone	Galaxy Note 4, iPhone 6s	1	Accelerom-eter, touch screen, gyroscope	Yes
N. Kim et al. [122]	16-24 kHz linear chirp signal	Amplitude of frequency	one speaker, one microphone	Samsung Galaxy Note 5	1	No	No
AGRS [123]	20 kHz continuous tone	Doppler shift	one speaker, one microphone	Mi3, Samsung S4, Sony M51w	1	gyroscope	Both
PatternListener [124]	18-20 kHz continuous wave	Phase	speakers, one microphone	Samsung C9 Pro, Huawei P9 Plus	1	motion sensors	No
AcouDigits [125]	19 kHz sin wave modulated signal	Frequency shift	one speaker, one microphone	Samsung Galaxy Note 5	1	No	Yes
EchoWrite [126]	20 kHz sinusoidal modulated audio signal	Doppler shift	one speaker, one microphone	Huawei Mate 9	1	No	Yes
C. Yiallourides et al. [127]	earpiece: 21 kHz; loudspeaker: 22.8 kHz; sinusoidal modulated audio signal	---	two speakers, two microphones	Samsung Galaxy S6	1	No	Yes
LLAP [128]	17-23 kHz CW sound signal	Phase	one speaker, two microphones	Samsung Galaxy S5	1	No	Yes
FingerIO [111]	18-20 kHz OFDM signal	Phase, Time of arrival	one speaker, two microphones	Samsung Galaxy S4	1	No	Both
Strata [129]	18-22 kHz BPSK signal	CIR, Phase	one speaker, two microphones	Samsung Galaxy S4	1	No	Yes
EchoTrack [130]	16-23 kHz chirp signal	ToF, Doppler shift	Two speakers, one microphone	Nexus 6P	1	No	Yes
BatTracker [132]	17 kHz	Doppler shift, Echo amplitude	one speaker, one microphone	Huawei P9, Samsung Note3	1	inertial sensor	No
SteerTrack [131]	20 kHz sinusoidal signal	ToA (Time of arrival)	one speaker, two microphones	Google Pixel, HTC U Ultra, Samsung Galaxy S6, LG G4, Huawei Mate8	1	No	Yes

result, we can achieve good identification accuracy. Therefore, the original sound signal generation plays a vital role in ultrasound sensing systems. According to the hand gesture recognition applications, many types of sound signals are adopted based on their characteristics, such as continuous wave (CW) signal [128], chirp signal [112], OFDM

signal [111], Zadoff-Chu (ZC) sequence [120]. All of them have advantages and limitations. CW signal is generally used to improve SNR; however, the spatial resolution needs to be improved. OFDM signal is adopted due to its low processing complexity and good synchronicity. Table 1 shows the sound signals used in human hand gesture recognition systems. From this table, we can observe that all these applications adopt 16 kHz - 24 kHz sound signal, which is inaudible to most people [133].

Besides, Table 1 also exhibits and compares the present systems in six aspects, and we will interpret them in the following. ''Extracted signal'' means that which signal will be extracted and analyzed to recognize hand gestures. The column ''Sensors'' indicates the number of speakers and microphones used in each system. ''Devices'' provides us with information about experimental devices. The ''Number of devices'' refers to how many devices work simultaneously to realize hand gesture recognition. And the ''Additional sensors'' shows if there are extra smartphone's built-in sensors used as auxiliary tools (e.g., accelerometer, gyroscope, compass). The selection of scheme depends on the aim of the system and can affect recognition accuracy. Therefore, before developing the system, we need to design system parameters and choose suitable hardware. Remarkably, we consider holding or touching the smartphone as a device-based pattern and placing the smartphone on the table or somewhere as the device-free pattern. This category can clarify the research contents of our paper.

2) SIGNAL ANALYSIS

Hand gestures performed by users would affect the ultrasonic signal propagation and change the sound signal waveform. Therefore, the echo signal is different from the original signal. We can identify hand gestures by analyzing the difference between the echo signal and the original signal. After capturing the sound signals, the changed signals are analyzed to detect hand gestures by using recognition algorithm. Generally, the hand gestures would change the frequency of the received signal, which is described as Doppler effect. And the hand movements also change the phase information and CIR observed at the received signal. Besides, we can localize the hand position by calculating the ToF. In this section, we introduce the basic recognition principles by using the echo signal. The types of extracted signals in human hand gesture recognition applications are shown as Table 1.

Doppler Effect: Christian Johann Doppler proposed Doppler effect in 1842. The main content of the Doppler effect is that the relative movement between source and observer will change the frequency of the signal received at the observer.

Specifically, when an object moving toward the audio source, the observed frequency will increase. Otherwise, the frequency will decrease if the object moves away from the sound source. SoundWave [87] confirms that waving the hand around the smartphone could lead to frequency changes, which can be used to detect hand motions.

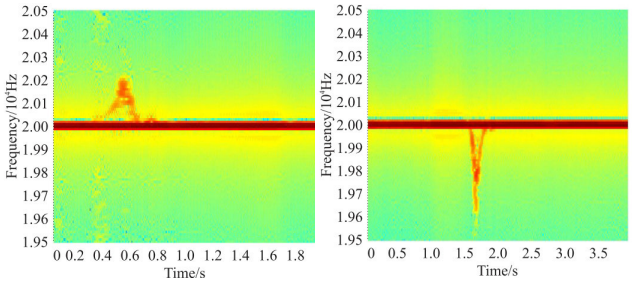


FIGURE 3. Time-frequency diagram of different hand motions [123]. (a) Moving towards the smartphone; (b) Moving away from smartphone.

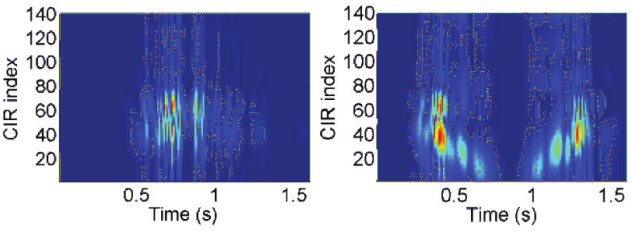


FIGURE 4. CIR information of different hand movements [118]. (a) Clockwise rotation; (b) Pull-push.

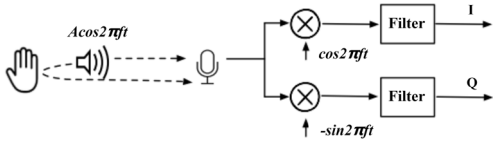


FIGURE 5. The basic I/Q demodulation structure.

Therefore, we can calculate the human hand speed by analyzing the frequency changes, described by Doppler shift. Fig. 3 depicts the time-frequency diagram of different hand motions. From Fig. 3, we can find that different hand movements lead to distinct frequency changes, which results in a specific pattern of the time-frequency diagram.

We can calculate hand velocity using the difference of the frequency between the received signal and the transmitted signal based on (1) and (2).

$$f' = \left(\frac{v + v_0}{v - v_0}\right) f \tag{1}$$

$$\Delta f = f' - f \tag{2}$$

where f' and f are the frequency of the received sound signal from microphone and the original signal from speaker, respectively; v and v_0 refer to the speed of sound in air and the velocity of the hand, respectively.

CIR Information: CIR describes the relationship between output and input. In other words, for linear and time-invariant (LTI) system, the output can be completely calculated using CIR. When user's hand moves around the smartphone, the CIR will change based on different gestures. Specifically,

various hand movements will produce distinct CIR information. Fig. 4 depicts CIR characteristics of different hand motions. From Fig. 4, we can find that different hand gestures correspond to specific patterns of the CIR information. Therefore, we can leverage the CIR information to recognize the hand gestures.

We can obtain the CIR information as follows according to [118]. Based on the theory of wireless communication, we can get (3).

$$R(n) = S(n) * h(n) \tag{3}$$

where $S(n)$ refers to the transmit signal; $R(n)$ is the received signal; $*$ represents the convolution calculation; $h(n)$ refers to the CIR information.

Then we can obtain $h(n)$ by using the Least Square (LS) equation which is expressed as (4) [118].

$$\begin{pmatrix} s_1 & s_2 & \cdots & s_L \\ s_2 & s_3 & \cdots & s_{L+1} \\ \vdots & \vdots & \ddots & \vdots \\ s_P & s_{P+1} & \cdots & s_{P+L-1} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_L \end{pmatrix} = \begin{pmatrix} r_{L+1} \\ r_{L+2} \\ \vdots \\ r_{L+P} \end{pmatrix} \tag{4}$$

where the matrix of s is the training matrix; vector h refers to the CIR information which is the data we intend to get; vector r is the received data; L is the crucial parameter to determine the channel impulse response length in discrete time; P can be calculated using $P + L = d$, where d is the length of data section in the training sequence.

Phase Information: The phase information is another feature to be utilized for hand gesture recognition. Since the sound signal will be reflected by the objects, the phase of the captured signal will reflect the propagation path changes. Therefore, it is usually utilized to track the hand trajectory and calculate the moving distance of user's hand. According to the presented hand gesture recognition systems in Section IV, the I/Q demodulation is a common method of extracting phase information from the received data. The received data are demodulated into In-phase (I) component and Quadrature (Q) component by I/Q demodulation algorithm. Specifically, the I component represents the component with the same direction of the captured signal while the Q component is orthogonal to the captured signal. And the basic demodulation structure is shown in Fig. 5.

We can obtain the I and Q components as follows according to Fig. 5. Supposing that the transmitted signal is $Acos(2\pi ft)$ and the signal would generate multiple path propagation, the captured signal from path p can be expressed as the following formula [128].

$$R_P(t) = 2A_P^{'} cos(2\pi ft - 2\pi f d_P(t)/c - \theta_P) \tag{5}$$

where $2A_P^{'}$ represents the amplitude of the captured sound signal; $d_P(t)$ refers to the time-varying path length; c and f are the speed of sound in air and the frequency of the emitted signal, respectively; θ_P is the initial phase lag caused by hardware delay.

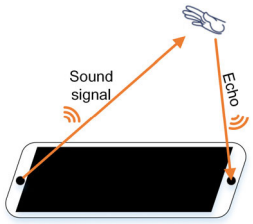


FIGURE 6. Localization using time of flight.

After obtaining the sound signals, they are multiplied with $cos(2\pi ft)$ and $-sin(2\pi ft)$ respectively. Then, we can get the I and Q components by using filters to eliminate the high frequency components. The I component and Q component are expressed as (6) and (7) [128].

$$I_P(t) = A_P^{'} cos(-2\pi f d_P(t)/c - \theta_P) \tag{6}$$

$$Q_P(t) = A_P^{'} sin(-2\pi f d_P(t)/c - \theta_P) \tag{7}$$

Finally, these two components are combined into a complex signal to obtain the phase information in path p.

$$\varphi_P(t) = -(\frac{2\pi f d_P(t)}{c} + \theta_P) \tag{8}$$

where $\varphi_P(t)$ and $d_P(t)$ represent the phase information of path p and the time-varying path length, respectively; c refers to the speed of sound in air; f means the frequency of the original signal; θ_P is the initial phase lag caused by hardware delay.

Time Information: Time information refers to the flight time of sound signals, such as time of arrival and time difference of arrival, etc. It is significant for distance measurement and localization due to its simplicity and precision. For human hand gesture recognition, we calculate the distance between hand and smartphone and realize the hand movement tracking using the time of flight. As shown in Fig. 6, the fundamental idea of this technique is that speakers emit the original sound signal and the microphones capture the echo signal reflected by hand motions. Then the time of flight between speakers and microphones could be used to calculate the distance between them. Finally, we can exploit the obtained distance information to locate hand position to achieve hand tracking. Assuming that the speaker transmits an ultrasonic signal at time t_0, the microphone receives the echo signal reflected by hand motions at time t_1, then the distance between speaker and microphone across hand can be expressed by formula (9).

$$D_{SHM} = (t_1 - t_0) \times c \tag{9}$$

where D_{SHM} represents the distance between speaker and microphone across hand; S, H, M represent the speaker, hand, and microphone respectively; c is the sound speed in air.

B. PREPROCESSING

Preprocessing algorithms are employed to obtain clean data and more effective information for feature extraction. Since there are plenty of noises contained in the sound signal captured by microphone, they need to be eliminated at first. Besides the environment noises such as other human behaviors and interference sound, the influence of signal drift stemming from the time elapses and device diversity also need to be removed. According to the characteristics of hand gesture recognition systems, we can adopt some effective processing approaches to eliminate interference and improve recognition accuracy.

Generally, filters can be used to remove ambient noise and reduce computation complexity. When we are interested in the data of a certain frequency range, we can utilize some filters and set thresholds to eliminate the high-frequency or low-frequency noises and keep favorite frequency range. Specifically, in hand posture recognition systems, the received signals from microphone contain plenty of environment noises (e.g., ambient sound signal, human activities, etc.) which can change signal waveform and reduce the quality of data. Thereby, to get valid data, the noises should be identified and then removed. Fortunately, we can use filters to achieve this goal. Many researchers have adopted various filters on their applications. Some researchers set threshold vector to remove environment noise [113], [123]. Besides, LLAP [128] uses a cascaded integrator comb (CIC) filter to enhance computational efficiency.

C. HAND GESTURE RECOGNITION METHODS

We divide the hand gesture recognition applications into two groups, including dynamic gestures and hand tracking based on different application requirements. The former usually serves as operation command and the latter can serve as data input. Since the dynamic gesture recognition can be considered as a classification problem, we can employ some machine learning and deep learning algorithms to identify hand postures. While hand tracking can be deemed as continuous and fine-grained localization. Therefore, the identification methods of tracking are very different from the gesture recognition. Rather than general machine learning algorithms, we usually exploit accurate mathematical models to calculate the trajectory of hand or to determine shapes or letters drawn by hand. In this part, we will concentrate on these recognition methods from two aspects, machine learning and deep learning for dynamic hand pose recognition and tracking methods for hand trajectory tracking.

Based on the dynamic hand gesture recognition applications in Table 2, we can apply various classification methods to recognize hand postures, e.g., HMM, SVM. Before using these classifiers, we must extract features from the original data to feed them into classifiers, such as frequency shift, amplitude, etc. From these presented applications, we notice that distinct systems generally extract specific features as the input data. For example, Dolphin [113] adopts manual gesture recognition and machine learning classifier. For manual classification, the system extracts F_t sequence (a weighted value of frequency shift, which denotes the frequency center changes caused by user at time t) of a complete gesture in chronological order as features. Then, the system employs support vector machine (SVM) and utilizes feature vector V (a value after interpolating for the ultrasonic data vector of a complete gesture) to train the classifier. UltraGesture [118] establishes a convolutional neural networks (CNN) model to identify hand gestures. It uses a complex CIR matrix calculated by LS algorithm as the features for CNN model training. As for hand tracking, it usually exploits geometric models rather than pattern-based methods. Therefore, these applications do not need the procedure of feature extraction.

After feature extraction, we obtain a feature vector that represents the input data. Next, we select classifier to recognize unknown gestures. Machine learning algorithms have been widely studied and adopted in various scenarios due to their satisfactory performance. For dynamic hand gesture recognition, we can exploit these algorithms to identify a specific action. As shown in Table 2, these systems adopt many machine learning methods including hidden markov model (HMM) and SVM to identify the different hand postures. Besides that, with the advance of deep learning algorithm, it has been widely adopted in various scenarios such as image processing, video retrieval, speech recognition, and natural language processing because it can discover and extract complex features automatically and achieve excellent recognition accuracy. For dynamic gesture recognition, we can convert sound data into an image and explore the capability of deep learning to classify the gestures. In this section, we concentrate on some common recognition methods, including some machine learning methods, deep learning methods, and tracking methods.

1) MACHINE LEARNING

Machine learning is usually used to solve classification problems. Since the gesture recognition can be considered as a classification problem, some systems utilize the machine learning methods to realize the hand gesture recognition. HMM and SVM are commonly used in hand gesture recognition applications. We introduce these methods as follows.

HMM: HMM is a kind of statistical analysis model and has become an essential method of signal processing. It describes a Markov process with unknown parameters and solves the problem based on time sequence and state sequence. Since the HMM can effectively process the time-varying sequence and achieves satisfying performance for the signal feature analysis, some researchers utilize the HMM model to recognize hand gestures. For example, AGRS [123] uses the HMM algorithm to realize the recognition of hand gestures. It first builds an HMM model for each kind of gesture and then calculates the similarity between the gesture data. Thus, we can establish HMM model to recognize the hand postures. Although the HMM classifier achieves better recognition accuracy, it takes a long time to train models and recognize gestures.

TABLE 2. Dynamic gestures recognition.

System	Preprocessing	Experimental scenarios	Behaviors	Recognition method	Accuracy	Findings
AudioGest [112]	Fast Fourier Transform (FFT) normalization, Audio signal segmentation	Living Room, Bus, Cafe, HDR Office, Lab	6 gestures, 5 users, 3900 samples	Direction: spectrogram analysis; Duration: direct measurement; Speed: using speed-ratio; Range: using range-ratio	95.1%	It could theoretically provide 162 control hand postures for applications by combining three waving factors.
Dolphin [113]	Noise elimination, Data normalization	Quiet environment, Outdoor, Noisy environment	24 gestures, 3 users,	Manual recognition, native bayes (NB), k-nearest neighbor classifier, bayes net (BN), SVM, Liblinear, random tree (RT),	94%	Combining manual and machine learning classification method can detect more gestures with high accuracy.
SonicOperator [117]	Noise elimination, Data normalization	Subway station, Restaurant, Indoor	24 gestures, 10 users, 36000 samples	RNN, SVM, Multi-layer perception (MLP), NB	95%	Five classification methods are adopted and the designed RNN network performs well and can improve the accuracy.
AGRS [123]	Noise reduction, Gyroscope calibration Data normalization	Quiet, Normal, and Noisy environments	8 gestures, 5 users, 2000 samples	HMM, SVM	95%	It uses gyroscope sensor auxiliary equipment and CFAR algorithm to reduce error recognition rate.
UltraGesture [118]	Down-conversion, Lowpass filter,	Different noise level, left hand, new user, with gloves on, etc.	12 gestures, 10 users, 12000 samples	CNN	>97%	Using CIR information and adding extra audio sensors can recognize fine-gained gestures and distinguish similar gestures.
AirLink [119]	Noise reduction	Laboratory	6 gestures, 11 users, 1260 samples	Mathematical calculation	96.8%	It enables multi-devices to share files and device pairing by using Doppler shift.
VSkin [120]	Cross-correlation, Extended Kalman Filter, Upsampling, Low-pass filter	Typical office and home environments	3 gestures, 10 users, 3200 samples	Estimation algorithm	Tapping events: 99.65%; Finger movem-ent s: 3.59 mm	It leverages the structure-borne sound and the air-borne sounds to achieve smartphone back surface tapping events and finger motions detection.
ForcePhone [121]	Calibration, Noise elimination, Normalization	Café	2 gestures, 27 users	Linear regression model, Setting thresholds	Touch force sensing: >97%; Squeeze phone body: 90%	It utilizes the structure-borne sound propagation property to measure the touch force and detect the predefined squeeze behavior.
N. Kim et al. [122]	---	An office environment	6 gestures, 10 users, 1800 samples	SVM	93%	It combines the structure-borne sound and the air-borne sound to recognize different types of hand grips.
PatternListener [124]	Coherent detection, Static components removal	Café, Office	130 patterns, 5 users	Pattern tree	>90% patterns within 5 attempts	It uses the phase information of the reflected signals to infer the unlock patterns.
AcouDigits [125]	Band-pass filter, Setting threshold	---	10 basic digits, 26 English letters, 10 users, 59600 samples	k-nearest neighbor classifier (KNN), SVM, artificial neural network (ANN)	Basic digits: 91.7%, English letters: 87.4%	It utilizes the frequency shift caused by hand movement to recognize the 10 basic digits and 26 English letters.
EchoWrite [126]	Median filter, Setting threshold, Gaussian filter Zero-one normalization, Binarization	Meeting room, Lab area, Resting zone	Stroke, Words; 6 users; 8640 samples	dynamic time warping (DTW), Bayesian language model	Entering texts at a speed of 7.5 WPM without practice and 16.6 WPM with practice	It leverages the Doppler shift caused by hand motion to recognize in-air text inputs without training and enables users to input texts at a comparable speed.
C. Yiallourides et al. [127]	Matched filter, Direct path removel	A quiet room, A noisy room, A noisy office	4 gestures, 9 users, 400 samples	SVM	77.5%	It uses two speakers to emit sound signal with different frequency and controls the start and end of recording signals by a computer.

SVM: SVM is a supervised learning model in the field of machine learning and is widely used in classification scenarios. It can effectively solve two-class classification and multi-class classification problems and has good classification performance. At the same time, SVM can transform the linear non-separable problem into a linear separable problem by extending the dimension of feature space. Therefore, it can be utilized to recognize hand gestures based on ultrasonic signal of smartphone (e.g., AGRS [123], Dolphin [113], SonicOperator [117]). However, it is difficult to be implemented when the sample size is large.

2) DEEP LEARNING

With the rapid development of computation capability, deep learning has attracted more attention in artificial intelligence area because it can discover many latent and complex features representing the original data. Deep learning originates from the study of artificial neural networks. It builds a neural network which simulates human brain to learn and analyze data. Especially, it can learn features from dataset automatically without the need for feature extraction. Since the methods of deep learning have high recognition accuracy and can learn features automatically, some teams adopt deep learning approaches to identify human hand gestures. These systems utilize the conventional model to recognize hand gestures, including recurrent neural network (RNN), CNN, etc.

RNN: RNN is a type of neural network for processing sequence data. Generally, it can be utilized to process the data with different sequence lengths. For example, some sequences (e.g., a continuous speech, continuous handwritten text) based on time are relatively long and their lengths are different. Moreover, they are difficult to be split into single samples for training deep neural networks (DNN) or CNN neural network because these sequences have time information. Fortunately, RNN provides a powerful capability to solve the problem of time sequence. In human hand gesture recognition systems, SonicOperator [117] considers that the hand gestures are comprised of many postures sequence in chronological order. Thus, it utilizes the RNN neural network to identify hand gestures due to the good performance of RNN in processing time series and classifying the sequential data.

CNN: CNN is the most common neural network model with deep structure and convolution computation. It uses the backward algorithm to train model parameters and SoftMax function to classify targets. Since CNN has favorite characteristics including local perception, weight sharing, and multi-convolution kernel, the computation cost is significantly reduced. Thereby, many researchers utilize CNN model to classify hand gestures. For example, UltraGesture [118], a system of human hand gesture recognition, establishes a CNN neural network to identify 12 hand postures. Its experimental results show that this system can recognize 12 postures with an average accuracy greater than 97%.

3) TRACKING METHODS

According to the hand tracking systems in Table 3, researchers usually employ some algorithms based on phase or establish some geometric models such as time-based model to track the hand movements. In this section, we concentrate on some commonly used methods for hand tracking in systems based on ultrasonic signal of smartphone.

Time-Based Model: Time information is usually utilized to track the hand trajectory. Specifically, the time of flight can be used to establish time model for hand trajectory tracking in systems based on smartphone-based ultrasound signal. According to the related studies, two models are usually established for calculating time difference. One is that two speakers emit ultrasound signal, one microphone receives the reflected signal to locate the hand [130]. The other is that one speaker emits ultrasound signal, two microphones receive the reflected signal, and the position of hand could be obtained by utilizing the time arrived at two microphones [131].

Phase-Based: Since the hand movements can influence the sound signal propagation and change phase of the signal, researchers leverage these changes to realize hand localization and tracking. Specifically, the phase changes of signal can be utilized to track the location of the hand because the changes effectively depict the change of hand position, which describes the path length changes. For 1-dimension, the hand movement distance can be estimated by transforming the phase change information into distance changes. For 2-dimension, the hand motion can be tracked by combining the movement path length changes and an initial position. For example, LLAP [128] first utilizes the phase changes to calculate the hand motion path length changes and then obtains an initial position based on the delay profile. It then achieves hand tracking by continuously updating hand location using these measurement results. Strata [129] calculates the distance changes of hand based on phase change information and estimates the absolute distance based on CIR information changes to obtain the initial position. It then combines the distance changes and the initial position to realize hand tracking.

IV. APPLICATIONS

With the development of technology and the improved performance of smartphone, quite a number of applications based on smartphone are emerging increasingly. Specifically, the hand gesture recognition systems based on ultrasonic signal of smartphone attract more attention due to their good performance, low deployment cost, and no-intrusive work pattern. In recent years, many human hand gesture recognition systems based on ultrasonic signal of smartphone have been proposed, bringing us with natural and novel methods of HCI.

In this section, we concentrate on the active sonar system based on smartphone. A smartphone can be treated as an active sonar system because it sends and receives ultrasonic signal. The received signal is changed by hand movement and can be exploited to recognize hand gesture

TABLE 3. Hand tracking.

Work	Preprocessing	Experimental scenarios	Behaviors	Accuracy	Dimension	Range	Mic/ speaker separation
LLAP [128]	I/Q demodulation, CIC filtering	Normal, Music, Speech, Speaker environments	26 letters, 11 words, 5 users	1D: 3.5 mm, 2D: 4.6 mm	1D, 2D	1D: 20 cm 2D: 10 cm × 10 cm	---
FingerIO [111]	Set distance threshold, Fine-tune distance estimate	Office	All kinds of shapes, 10 participants	Median error: 8 mm	2D	10 cm × 10 cm	13.5 cm
Strata [129]	Bandpass filtering, Frame detection	Student office	Diamond, Triangle, Circle, 5 users	0.3 cm distance tracking error, 1 cm 2D tracking error, 0.6 cm 2D drawing error	1D, 2D	1D: 40 cm 2D: 31.1 cm	14 cm
EchoTrack [130]	Bandpass filtering, Multipath elimination	Laboratory	Straight line, Triangles, 20 users	76% within 3 cm error, 48% within 2 cm error	3D	80 cm	14 cm
Battracker [132]	Bandpass filtering	Cluttered Laboratory, Typical bedroom	Diamond, Triangle, Word, etc.	3D: 90% within 1cm error 2D: < 1 cm	3D, 2D	3 m × 3 × 3 m	---
SteerTrack [131]	Eliminating symbol time offset	Real driving environments: local road and highway	Track the rotation angle of steering wheel, 5 volunteers	4.61 degrees error	3D	---	---

or trajectory. We first review the state-of-the-art applications of hand gesture recognition based on ultrasonic signal of smartphone. Then, we divide them into two groups: dynamic gesture recognition and hand tracking. The dynamic gesture recognition systems identify specific hand postures, such as flick, push, and pull, etc. We introduce these systems from preprocessing techniques, experimental scenarios, recognized behaviors, participants, samples, recognition methods, their findings, and performance (see in Table 2). The hand tracking systems refer to hand trajectory tracking, such as writing letters and drawing circles or rectangles, etc. We analyze these systems in several aspects, including preprocessing techniques, experimental scenarios, recognized behaviors, participants, samples, performance, dimension, moving range, and the distance between speaker and microphone (see in Table 3).

A. DYNAMIC GESTURE

Currently, hand gesture recognition using ultrasonic signal of smartphone has drawn more attention and quite a number of studies are continuously emerging because they bring a new way to interact with computer and improve the quality of HCI. In this section, we concentrate on some dynamic gesture recognition applications, introduce their main idea and major processing work, analyze the performance of these systems, and point out some useful findings, as shown in Table 2. Then, we present future research directions.

1) AUDIOGEST [112]

In 2016, W. Ruan *et al.* propose a device-free and training-free hand gesture recognition system based on ultrasonic signal of smartphone, called AudioGest [112]. The fundamental idea of this system is that Doppler shift can be calculated based on the received ultrasonic signal and the speed of hand can be deduced. In AudioGest, authors transmit 19 kHz audio signal and adopt denoising pipeline to suppress signal drifting and extract the weak echo signal. In the test phase, extensive experiments are conducted to evaluate the performance of this system. Three distinct mobile devices including laptop, tablet, and smartphone are deployed to evaluate this system with 5 participants performing 6 hand gestures in 5 environments.

The experiment at lab shows that AudioGest can recognize 6 gestures with an average accuracy of 94.15%. What's more, the authors verify the influence of different factors on the classification accuracy, such as the orientation angle of device and the distance between device and hand. AudioGest also evaluates system performance in other four environments. The experimental results prove that this system exhibits good robustness to real-world places. Especially, AudioGest can estimate many hand movement states including hand moving directions by using audio spectrogram, hand moving time in air (long, normal, short) by using direct time interval measurement, hand moving speed (slow, medium, fast) by using speed-ratio and waving range (wide, middle, small) by using range-ratio with high accuracy. Based on the combination of these states, authors argue this system can achieve more fine-grained hand gesture recognition by considering various factors. Theoretically, AudioGest could provide $6 \times 3 \times 3 \times 3 = 162$ control hand postures for applications by combing these waving factors (see in Fig. 7).

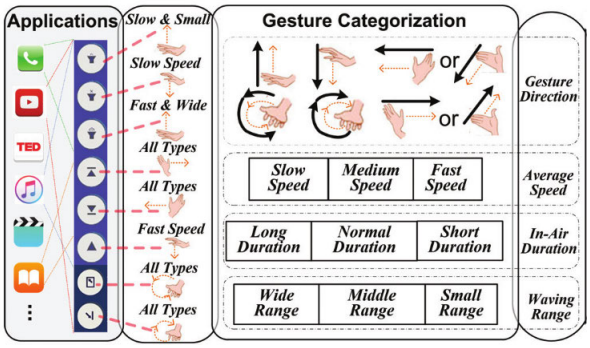


FIGURE 7. The overview of hand gestures with waving factors [112].

In their future work, they will consider extracting more features from the spectrogram to recognize more hand postures. Many other methods can be applied to address the influence of environmental movements, such as combining other smartphone's built-in sensors, or emitting modulated sound signal (e.g., multiple frequency shift keying sound signal).

2) DOLPHIN [113]

In 2014, Q. Yang *et al.* present an in-air hand gesture recognition system called Dolphin [113]. This system uses the speaker and microphone built in smartphone to emit and receive ultrasonic signal. Specifically, the speakers transmit a 21 kHz continuous tone and the microphone captures the echo signal reflected by hand movements. Then this system extracts the Doppler shift information caused by hand motions from the echo signal and utilizes the Doppler shift information to identify hand gestures. Different from other applications, the authors propose a method that combining the manual recognition algorithm and machine learning algorithm to classify more hand postures. To be specific, Dolphin predefines many gesture groups. When recognizing a gesture, it first divides the gesture into a corresponding gesture group using the proposed manual recognition method, and then classifies the gesture into a finer gesture label by conducting machine learning algorithms to the group.

In the experimental phase, two smartphones and one tablet are deployed to test the performance of Dolphin with 3 participants. The authors first use manual recognition approaches to classify the hand postures into 10 groups and then employ 7 machine learning algorithms to recognize the similar hand gestures and compare the algorithm performance. The experimental results show that Dolphin can recognize 24 predefined hand gestures with an average accuracy of 94%. Furthermore, the authors evaluate Dolphin by establishing an Android plugin (see in Fig. 8(a)) and designing two games (see in Fig. 8(b)), and the results prove that Dolphin can perform game control accurately. The future work of Dolphin will focus on improving the performance in some complicated scenarios and optimizing the energy consumption of the smartphone.

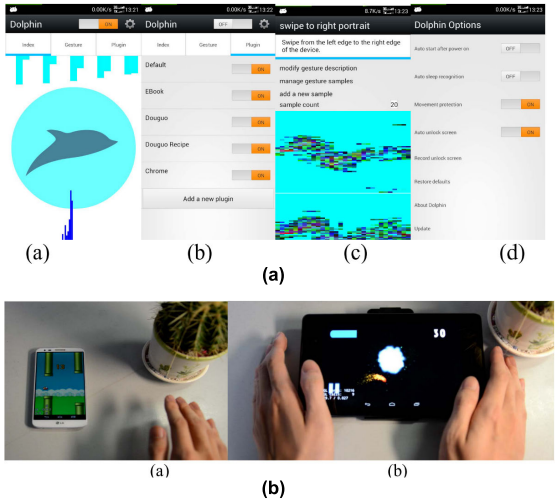


FIGURE 8. (a) Android plugin interface; (b) Two games developed by using Dolphin [113].

3) SONICOPERATOR [117]

In 2017, X. Li *et al.* present an in-air hand gesture recognition system called SonicOperator [117]. This system utilizes the speaker and microphone built in smartphone as the transceiver and analyzes the Doppler shift caused by hand motions to recognize multiple hand gestures. The speaker first emits a 21 kHz ultrasonic signal and the microphone captures the reflected signal. Then, the captured sound signal in time domain is transformed into frequency domain by conducting FFT. After eliminating noise, normalization is performed to keep the data into same scale. The authors deem that the gestures are comprised of a series postures in chronological order; therefore, they choose RNN network as the classifier to recognize hand gestures and use the transfer learning algorithm to transfer the knowledge of feedforward neural network to the RNN.

In the experimental phase, four different smartphones are used to evaluate system performance with 10 participants. Specifically, the authors collect a total of 36000 samples in three environments and compare RNN with other four machine learning methods (see in Fig. 9). Extensive experiment results show that SonicOperator can increase identification precision and recognize 24 pre-defined gestures with an average accuracy of 95%. Furthermore, the experiments demonstrate that the proposed RNN method has good performance when the number of hand posture types increases, while the other methods obtain bad recognition accuracy.

4) AGRS [123]

In 2018, Z. Xu *et al.* propose a gesture recognition system called AGRS [123]. This system utilizes smartphone to implement gesture recognition. The principle of the system can be described as follows. The speaker built in smartphone transmits a 20 kHz ultrasonic signal and the microphone in the same mobile phone receives the echo signal reflected by hand movements. And this system extracts the Doppler shift from echo signals as the recognition information to

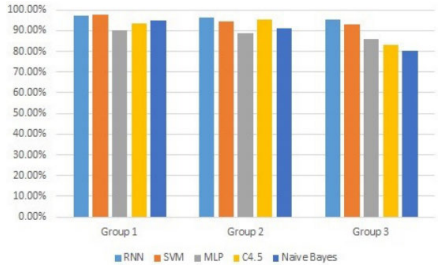


FIGURE 9. The comparison results of five classifiers [117].

Environment	Gyroscope	Position	Gestures								Average accuracy
			←	→	↑	↓	↘	↖	↗	↙	
Quite	Yes	Table	97.27	98.45	99.63	99.73	97.09	97.90	97.54	97.72	98.17
		Hand	96.35	97.67	99.34	99.16	96.89	96.75	96.87	96.53	97.45
	No	Table	96.95	98.13	99.21	98.68	96.95	96.54	96.75	96.76	97.50
		Hand	95.24	96.36	98.66	98.24	95.43	94.65	94.12	93.84	95.82
Normal	Yes	Table	95.12	96.54	99.21	99.54	96.25	95.38	96.64	94.32	96.78
		Hand	93.67	94.84	98.79	98.36	93.29	93.97	93.23	92.56	95.20
	No	Table	94.89	95.48	98.45	98.67	95.56	94.68	95.45	93.76	95.87
		Hand	90.21	90.46	96.35	96.28	90.86	90.45	90.31	69.34	89.28
Noisy	Yes	Table	68.56	70.23	84.35	81.35	71.33	72.64	73.13	69.27	73.86
		Hand	65.45	63.98	73.62	70.43	69.87	70.62	69.98	66.48	68.80
	No	Table	67.54	68.31	83.86	81.21	70.45	71.32	71.84	68.46	72.87
		Hand	58.67	60.23	68.52	65.47	64.76	65.12	66.57	64.21	64.19

FIGURE 10. The average recognition accuracy of hand gestures in three environments (%) [123].

identify hand gestures. In addition, the authors leverage the gyroscope sensor built in smartphone as an auxiliary tool to detect the placement state of the phone and employ the constant false-alarm rate (CFAR) algorithm to reduce the error recognition rate. Specifically, the FFT algorithm is first conducted to transform the captured sound signals into frequency domain and then the noise removal method is deployed to eliminate the noise interference. After extracting the frequency change information and performing data normalization algorithm, HMM and SVM classifiers are utilized to classify hand gesture samples.

In the experimental phase, the authors test the performance of AGRS on three smartphones with 5 volunteers performing 8 kinds of hand gestures. They collect a total of 2000 samples in 3 environments, and the experimental results are shown in Fig. 10. Experiments show that AGRS achieves a recognition accuracy more than 95% in quiet and normal environments. The recognition accuracy is severely influenced by environmental noise. The future work of AGRS will concentrate on the system power consumption reduction and the classification precision improvement under the noisy scenarios.

5) ULTRAGESTURE [118]

In 2018, K. Ling *et al.* present UltraGesture [118], a gesture recognition system using more microphones in smartphone. The characteristics of this system are that it uses the CIR information instead of Doppler shift to recognize gestures and exploits more microphones to resolve the problem that similar gestures are difficult to be classified. The authors use the LS estimation algorithm to calculate CIR after performing down conversion and low pass filtering. Then they conduct a differential operation to obtain the dCIR information and take

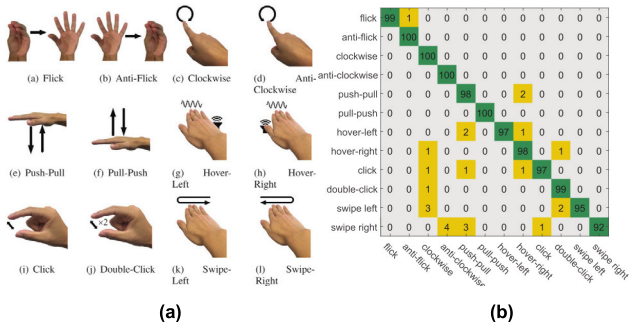


FIGURE 11. (a) 12 types of hand gestures; (b) The confusion matrix of using smartphone only [118].

the dCIR information as the input data due to the obvious feature of gesture in dCIR image. Finally, they design a CNN model to recognize finger motions.

In the experimental phase, the authors use a smartphone and a designed speaker-microphone kit to conduct their experiments. They require 10 participants performing 12 types of gestures (see in Fig. 11(a)) under some different environments. The experimental result (see in Fig. 11(b)) by using only speaker and microphone built in smartphone shows that UltraGesture can achieve an average accuracy up to 97.92%. The authors also use an additional speaker-microphone kit to assess the influence of microphone number. The experimental accuracies are 92.75%, 95.00%, 96.83%, and 98.58% for 1 to 4 microphones, respectively, which proves that the increase of microphone number could improve the recognition accuracy and help to classify similar gestures.

6) AIRLINK [119]

In 2014, K.-Y. Chen *et al.* [119] propose a device-free system called AirLink. This system utilizes the Doppler shift caused by hand movement to recognize gestures and then shares files between devices. AirLink requires each smartphone in the experiment to transmit the ultrasound signal at 18.8 kHz, and the microphone built in smartphone captures the echo signals. Each smartphone analyzes the echo signal and detects the hand gestures. Then the smartphones send the detection results to the central server to enable files sharing. Besides that, this system can identify the hand motion direction by combining the detection results of all the smartphones. For example, T and A refer to toward/away from the smartphone respectively, and X represents toward and then away from the smartphone. Thus, the M2R gesture can be represented as A-A-T under a three-smartphone scenario as shown in Fig. 12.

In the experimental phase, two different smartphones are used to evaluate the system in a laboratory scenario. Specifically, 11 persons perform 6 hand gestures in a three-smartphone environment. The experimental results show that the average accuracy is up to 96.8%. In addition to sharing files between devices, AirLink can also enable device pairing. In future work, this system can decrease the impact

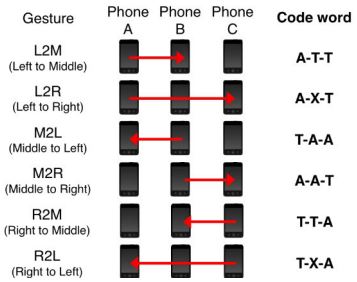
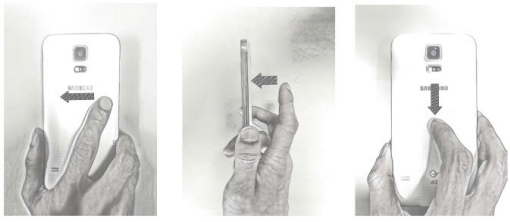


FIGURE 12. 6 gestures and their code words under a three-smartphone scenario [119].



(a) Back-Swiping (b) Back-Tapping (c) Back-Scrolling

FIGURE 13. Gestures performed on the back of the smartphone [120].

of the nearby hand motions and reduce the energy consumption of the mobile phone by configuring the smartphones to transmit the ultrasound signal only in specific scenarios.

7) VSKIN [120]

In 2018, K. Sun *et al.* present a novel system called VSkin [120], which can achieve fine-grained tapping events and finger movement detection based on ultrasonic signal of smartphone. Especially, these two gestures are performed on the back surface of the mobile phone, as shown in Fig. 13. The main idea of VSkin is that the speaker built in smartphone emits a 17 kHz - 23 kHz narrow inaudible sound signal and the microphones embedded in smartphone capture the echo signals mixed with two types of sound signals, including the structure-borne sound signal (the sound which travels through the smartphone body structure) and the air-borne sound signals (the sounds reflected by hand, wall, etc.). After analyzing the received sound signals, the authors first separate these sound signals according to their propagation path and then measure the amplitude and phase of each path. For finger movement calculation, they extract phase information from the air-borne path sound signal reflected by the hand and establish a finger movement model to measure the path length. Specifically, the authors utilize the change of path length to estimate the finger movement distance and achieve swiping and scrolling gesture recognition. For finger touch sensing, the authors use the structure-borne sound signal and extract the delay samples and magnitude of differential impulse response values to detect finger tapping.

In the experimental phase, extensive experiments are carried out to evaluate the performance of VSkin. Four different smartphones are utilized to conduct the tests with

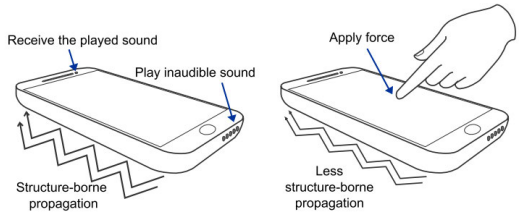


FIGURE 14. The structure-borne propagation with and without an applied force [121].

10 students performing 3 gestures (see in Fig. 13) in typical office and home environments. The results show that VSkin can measure the finger movement distance (finger moves 6 cm) with an average error of 3.59 mm, detect finger tapping events with an accuracy of 99.64%, and recognize swiping gesture with an accuracy up to 94.5%. Moreover, VSkin can achieve a low latency of 4.83 ms when performing gestures on the smartphone. VSkin leverages one speaker and two microphones to recognize hand gesture. One microphone is close to the speaker and another one is at the opposite side of the speaker. The implementation leads to this algorithm unavailable for other smartphones because the positions of microphones and speakers of the smartphones may be very different from VSkin. The next work of VSkin is to improve the algorithm to accommodate different types of smartphones with their own speaker/microphone layout.

8) FORCEPHONE [121]

In 2016, Y.-C. Tung *et al.* propose ForcePhone [121], which can realize the force measurement applied to the smartphone touch screen and identify the squeeze applied to the smartphone body. Similarly, it turns the phone into an active sonar that the speaker of smartphone emits an inaudible sound signal and the microphone on the same phone captures the played sound signal. ForcePhone leverages the structure-borne propagation property to measure the force. When applying force to the touch screen, the restricted phone body will degrade the sound signal traveling through phone body (see in Fig. 14) and the changed signal can be utilized to estimate the force. Meanwhile, ForcePhone uses the accelerometer and gyroscope readings to remove the other noises caused by movements. It establishes a linear regression model to estimate the applied force. Besides force measurement, this system can also detect the squeeze applied to the phone body. However, it only recognizes predefined squeeze behavior.

Extensive experiments are conducted to test the performance of ForcePhone. One android smartphone and one iOS smartphone are used in experiments with 27 users performing touching and squeezing gestures. The experimental results show that ForcePhone can achieve two different levels of touch force measurement with an accuracy of 97% and detect the squeeze of the phone body with an accuracy higher than 90%. Furthermore, the authors design ForcePhone-based apps and the users are satisfied with these apps and think they are useful and helpful. In future work, ForcePhone will focus

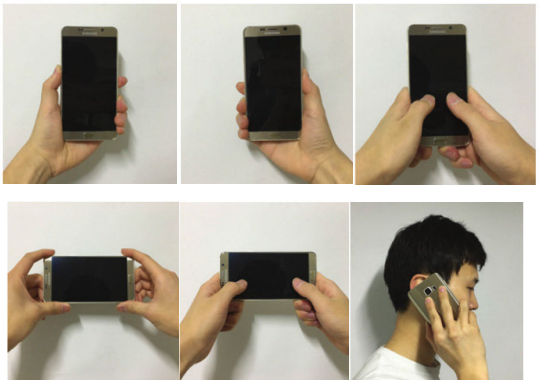


FIGURE 15. Six types of grip gestures [122].

on being implemented in some smaller and wearable devices, such as smartwatch. Therefore, this approach can expand its research areas and be applied to more applications.

9) GRIP SENSING FOR SMARTPHONE [122]

In 2017, Kim, N. *et al.* present a novel ultrasonic sensing system based on the smartphone. This system can detect how human grasps the smartphone. It first uses the speaker built in the smartphone to transmit a linear chirp signal with the frequency range 16 kHz - 24 kHz, then utilizes the microphone embedded in the same smartphone to capture the echoes. Since the echoes comprise the structure-borne sound (which travels through the smartphone body) and the air-borne sound (which travels through the air), the mixed echoes' spectrum of each grip is unique and it can be leveraged to recognize the grips. After receiving the sound signal from the microphone, the authors perform FFT algorithm to obtain the frequency-domain data. Afterward, they extract 172 features from the processed frequency-domain data and use them to train SVM classifier to recognize grips.

In the experimental phase, one smartphone is used to evaluate the performance of this system in an office scenario and 10 users are required to perform 6 types of grips as shown in Fig. 15. The experiments demonstrate that this system can achieve identifying 6 types of grips with an average accuracy of 93%. The future work of this system will concentrate on improving the recognition accuracy and adapting itself to various practical environments.

10) PATTERNLISTENER [124]

In 2018, M. Zhou *et al.* propose PatternListener [124], a novel active sonar system that can infer the unlock pattern of the smartphone. As shown in Fig. 16, there are some unlock patterns with different number of lines. This system first leverages the speaker built in the smartphone to transmit an imperceptible audio with a frequency ranging from 18 kHz - 20 kHz and the microphone on the same phone to record the signals reflected by fingers. Then, this system demodulates the recorded sound signals by using the coherent detection and utilizes a local extreme value detection (LEVD) algorithm and linear interpolation algorithm to eliminate the

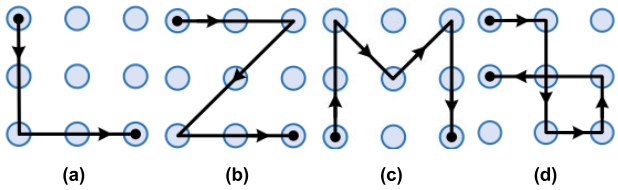


FIGURE 16. Example patterns with different lines [124]. (a) Two lines; (b) Three lines; (c)Four lines; (d)Five lines.

noises (e.g., signals reflected by the wall or other objects). Afterward, PatternListener uses a designed turning points identification (TPI) algorithm to segment the preprocessed signal into fragments corresponding to each line. At last, it uses the phase changes of the reflected signal to calculate the path change length and establishes a pattern tree to infer the unlock pattern.

In the experimental phase, two different types of android smartphones are adopted to evaluate the performance of PatternListener in 2 environments and 5 participants are required to draw the unlock pattern which they chose from the 130 predefined patterns. Specifically, various effect is considered when the authors assess PatternListener, including the pattern complexity, unlock gestures, drawing speed, surrounding objects, types of smartphones, and ambient noise. Extensive experimental results show that PatternListener can infer more than 90% unlock patterns within 5 attempts. Furthermore, the authors propose two methods to defend against PatternListener's attack, including prohibiting the usage of microphone in the background and randomizing the layout of the pattern grids.

11) ACOUDIGITS [125]

In 2019, Y. Zou *et al.* design a novel system called AcouDigits [125], which can recognize the basic digits and English alphabets using the frequency shift caused by hand movement. This system adopts a sin modulated sound signal with a frequency of 19 kHz emitted by the smartphone's embedded-in speaker. After recording the reflected sound, the authors use a band-pass filter to eliminate the noise and enhance SNR and transform the time domain data into frequency domain data to analyze the frequency shift. They extract 5 time domain features and 4 frequency domain features to train the classifiers including KNN, SVM, and ANN. For KNN model, they directly feed the 9 features into KNN model to train the KNN classifier. For SVM and ANN models, they further extract 5 statistical features (e.g., range, mean value, variance) from the 9 extracted feature data as the input data to train these classifiers.

In the experimental phase, one android smartphone is applied to test the AcouDigits with 10 volunteers and the experiments last for 10 days. The authors collect 44000 samples of the basic digits and 15600 samples of the English letters. To be specific, they utilize SVM, ANN, and KNN to classify the basic digits, respectively, and leverage ANN to recognize the English letters. Comprehensive experimental results demonstrate that AcouDigits can recognize 10 basic

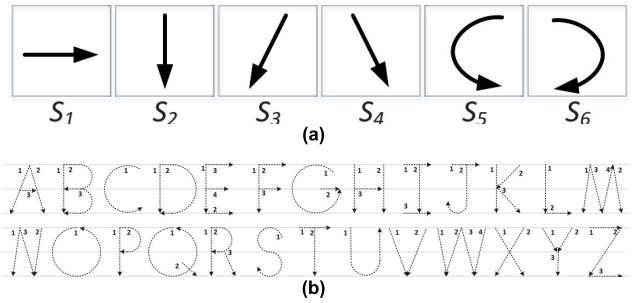


FIGURE 17. (a)The six fundamental strokes of English alphabets; (b) The English alphabets' stroke order [126].

digits with an average accuracy of 91.7% and identify 26 English letters with an average accuracy up to 87.4%. The authors also use a CNN model to assess AcouDigits and the results show that AcouDigits can recognize the digits and the letters with an accuracy of 94.9%.

12) ECHOWRITE [126]

In 2019, Y. Zou *et al.* present a training-free text-input system called EchoWrite [126], which can recognize the in-air finger writing based on ultrasonic signal of smartphone. This system employs the speaker built in smartphone to emit a 20 kHz acoustic signal and the microphone to capture the echoes. After obtaining the echoes, EchoWrite first transforms the time domain data into frequency domain data by performing short-time Fourier transform (STFT). Then, it uses a median filter and sets an energy threshold to eliminate noises. After that, EchoWrite smooths the spectrogram by a Gaussian filter and conducts normalization and binarization algorithms to the smoothed data to get a clean spectrogram which depicts the Doppler shift. At last, EchoWrite extracts Doppler shift features, utilizes DTW to identify the strokes, and leverages the Bayesian language model to realize input text inference.

In the experimental phase, the authors employ an android smartphone to evaluate the performance of EchoWrite with 6 users in 3 environments. The participants are required to perform 6 types of strokes (the strokes come from the 26 uppercase English alphabets, as shown in Fig. 17) and write 10 words. The experimental results show that EchoWrite can achieve stroke recognition with average accuracies of 94.4%, 94.9%, 93.2% in the three scenarios, respectively and can infer the input words with an accuracy of 94.9%. Moreover, this system enables participants to input texts at a speed of 7.5 words per minute (WPM) without practice, and 16.6 WPM after approximately 30-minute practice. In the future work, EchoWrite will focus on the following aspects, including availability of wearable devices, improvement of its robustness to some burst noises (e.g., knocking the table), and redefinition of their own input gestures.

13) C. YIALLOURIDES *et al.* [127]

In 2019, C. Yiallourides *et al.* propose a novel hand gesture identification method. This system uses the two speakers built in smartphone to transmit ultrasound signal with different frequency and utilizes two microphones embedded in the same phone to record the echo signal. After obtaining the echo signal, it removes the uninterested parts by a matched filter and direct path (the sound signal propagates directly from the speaker to microphone) removal algorithm. Then, it estimates the signal to noise ratio and extracts four statistical moments as the features. Since this system leverages two microphones to collect the reflected signal, an 8-dimensional feature vector will be obtained for each gesture. At last, the feature vectors will be employed to train the SVM classifier to recognize gestures.

In the experimental phase, an android mobile phone is used to evaluate the performance of this system. The authors recruit 9 participants to perform 4 gestures in three environments and collect a total of 400 samples (360 gestures and 40 noise observations). Moreover, they use a computer to start and end the sound signal recording. Comprehensive experiment results demonstrate that this system can achieve an average accuracy of 77.5% when choosing radial basis function (RBF) kernel SVM and setting the parameters $C = 70$, $\gamma = 0.01$.

B. HAND TRACKING

Recently, research about hand tracking using ultrasonic signal from smartphone is drawing more and more interest. These systems focus on hand motions, e.g., drawing a shape or writing a letter. The drawing can be utilized as effective information input and expand the way of HCI. Specifically, we can recognize hand writing and drawing by this tracking technique. There are already many hand tracking systems by using the ultrasonic signals of the smartphone, as shown in Table 3. In this section, we introduce these systems, analyze the main contents and the performance of these systems, and point out some remaining work for future studies.

1) LLAP [128]

In 2016, a hand trajectory tracking system based on ultrasonic signal of smartphone, called LLAP, is proposed by W. Wang *et al.* This system leverages the speakers to emit a 17 kHz - 23 kHz CW signal and the microphones to capture the sound signals reflected by hand movements. The authors first collect the sound data and extract the phase information because the phase changes can identify fine distance variation compared with Doppler shift. Concretely, they use I/Q demodulation to obtain the complex signal and then separate it into static vector and dynamic vector. The former stems from LOS path or static objects and the later comes from the hand movements. Then, they convert the phase information into distance information according to the dynamic vector to achieve one-dimensional distance measurement (see in Fig. 18(a)). Furthermore, they achieve 2D hand gesture tracking (e.g., hand drawing shapes or words as shown in Fig. 18(b)) by combining the fine-grained phase and coarse-grained delay measurements.

Two smartphones based on Android and iOS system are used to evaluate LLAP with 5 participants in 4 scenarios.

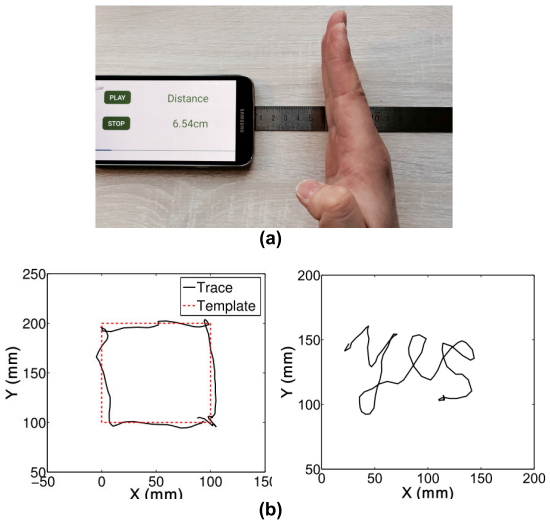


FIGURE 18. Applications of LLAP [128]. (a) Distance measurement; (b) Hand tracking for drawing shapes and words.

Results show that this system achieves hand movement distance measurement and hand gesture tracking with higher accuracy, lower latency, and much higher speed. Specifically, LLAP evaluates the system performance using different criteria for 1D and 2D tracking. It obtains mean movement distance error of 3.5 mm when the hand moves 10 cm at a distance of 20 cm and can reliably measure distance with velocity from 4 cm/s to 25 cm/s. It is robust to ambient noises and obtains a mean movement distance error of 5.81 mm. Besides, it achieves a tracking error of 4.57 mm, 26 Latin character recognition accuracy of 92.3%, and 11 words (e.g., yes, can) recognition accuracy of 91.2%, respectively. For system responsiveness, its latency is less than 15 ms on the implementation of smartphone. However, LLAP can only realize tracking a single object. Thereby, it regards the fingers and the hand as an integrated target and cannot identify multi-finger movement, e.g., ''pinch''. One of the future works of LLAP will be distinguishing the multiple fingers to realize multiple objects tracking by using more microphones.

2) FINGERIO [111]

In 2016, R. Nandakumar *et al.* present FingerIO [111], a device-free hand tracking system that can achieve millimeter-level tracking accuracy. The main idea of this system is to turn the smartphone into an active sonar system by treating the speaker and microphone in smartphone as sound transceivers. The speaker first emits an 18 kHz - 20 kHz OFDM signal with cyclic suffixes. The cyclic suffixes can be utilized for correcting sampling errors to achieve fine-grained hand tracking. Then the microphones capture the echo signals reflected by hand motions. The echo signals from two microphones will be analyzed to compute the distances between hand and microphones to realize 2D tracking.

In the experimental phase, a phone with two microphones (13.5 cm distance between two microphones) is deployed to evaluate FingerIO with 10 users in the office. The users are

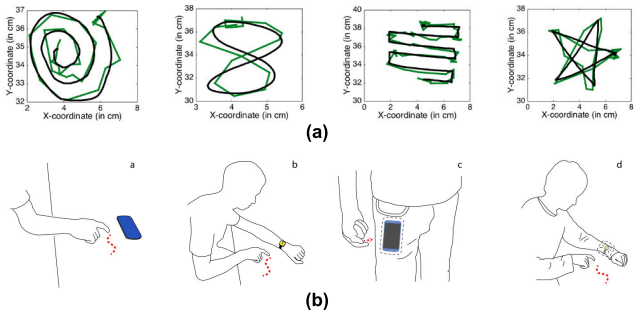


FIGURE 19. The tracking results and application scenarios [111]. (a) The tracking accuracy, the black lines refer to the ground truth trace while the green lines represent the trace that FingerIO tracks; (b) The application scenarios with and without covering.

required to draw any shapes in a range of 10 cm × 10 cm area. Extensive experimental results show that FingerIO can achieve 2D hand tracking with an average 8 mm error and the recognition accuracy severely decreases when interference occurs within 50 cm from the phone. Furthermore, the authors design a smartwatch to extend the interaction area to a range of 0.5 m × 0.25 m, and achieve 2D tracking with a mean error of 1.2 cm. Fig. 19 shows some tracking results and application scenarios. It can work well even when the device is in pockets. This benefit extends its application area. The future work of FingerIO will focus on realizing 3D tracking by using three microphones, tracking multiple fingers to detect gestures (e.g., zoom out, pitch), reducing the energy consumption of the smartphone, and tracking object with a moving device.

3) STRATA [129]

In 2017, S. Yun *et al.* propose a device-free hand tracking scheme based on ultrasonic signal of smartphone called Strata [129]. This system takes the multipath effect into consideration; thereby, it estimates the CIR information to track the hand trajectory. Strata first leverages the speaker to transmit an 18 kHz - 22 kHz inaudible sound signal and then uses the microphones to receive the reflected sound signal. After estimating the CIR information from the collected sound signal, the relative distance changes and the absolute distance information calculated using CIR are combined to track the hand movements.

Extensive experiments are conducted to evaluate the performance of this system. A smartphone with one speaker and two microphones is deployed to test Strata with 5 participants in a student office. The experimental results (see in Fig. 20) show that Strata achieves a median error of 1.0 cm for relative distance measurement under 1D tracking. To be specific, 20 cm, 30 cm, 40 cm are the initial distance between finger and the microphone and the finger moves 10 cm from these initial positions towards the microphone, respectively. And Strata obtains a median error of 1.01 cm for 2D tracking. Besides, it has favorite robustness and increases 1 mm median error under music background. For shape recognition, it obtains a median error of 0.57 cm for different shapes from

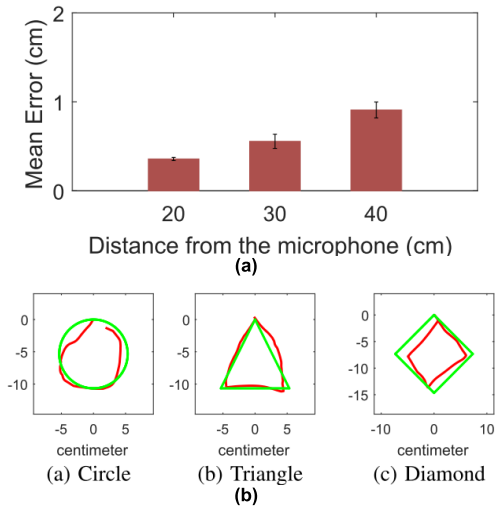


FIGURE 20. Tracking accuracy [129]. (a) 1D tracking accuracy; (b) Tracking accuracy of hand drawing shapes in 2D space.

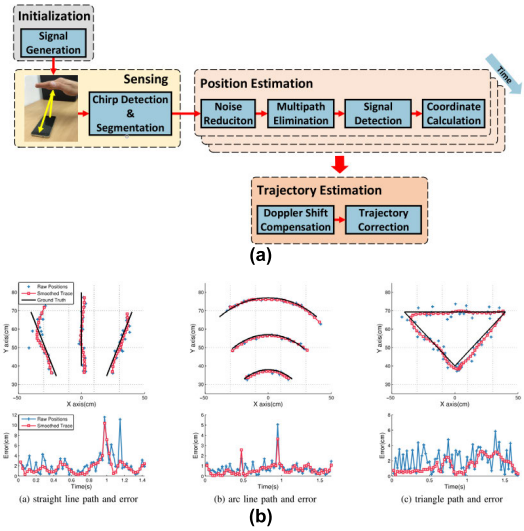


FIGURE 21. (a)The workflow of EchoTrack; (b)Path tracking results [130].

5 users. Its latency is 12.5 mm because it processes the signal for each this cycle. One of the future work of Strata is further improving the tracking accuracy.

4) ECHOTRACK [130]

In 2017, H. Chen *et al.* present EchoTrack [130], a hand tracking system based on ultrasound signal of smartphone. The authors leverage the speakers to emit a 16 kHz - 23 kHz chirp sound signal and the microphone to capture the echo signal reflected by hand movements. They utilize the ToF to measure the distance between hand and microphones to realize hand localization and tracking by continually localizing hand. Moreover, the authors leverage the Doppler shift compensation and trajectory correction algorithms to improve accuracy. The workflow is shown in Fig. 21(a). From this architecture we can get that there are four processing parts to realize hand tracking, including phase initialization, sensing phase, position estimation phase, and trajectory estimation.

In the experimental phase, a smartphone with two speakers and a wood board which has a similar size with the hand are used to test the performance in a laboratory. The smartphone continuously locates the wood board to realize the tracking, as shown in Fig. 21(b). Extensive experiment results show that EchoTrack can achieve trajectory tracking with an accuracy of 76% within 3 cm locating error and 48% within 2 cm locating error. Especially, the experiments validate the probability to tracking the in-air hand movement using smartphone-based ultrasonic signal. The future work of EchoTrack will concentrate on tracking in multi-user scenarios and improving energy efficiency by designing an effective dynamic scheduling mechanism.

5) BATTRACKER [132]

In 2017, B. Zhou *et al.* propose an infrastructure-free object location tracking system in 3D indoor space using inertial and acoustic data, called BatTracker [132]. The speaker emits a sound signal pulse of 17 kHz with 1 ms. BatTracker exploits echoes reflecting from the object and utilizes distance measurements to mitigate error accumulation. It eliminates the noise from multi-path reflection and complex environment and combines Doppler effect and echo amplitude to build the relationship between echoes and objects. After finding the initial position of an object according to reference object location, this system updates the track along with time according to inertial and acoustic sensors. The inertial data are used for object position prediction and acoustic data are used for position correction, which is implemented by motion model. For more noise measurements, BatTracker utilizes direction observation from position changes to suppress them, which is achieved by the observation model. Then, it applies a probabilistic algorithm to calculate the continuous location of the object. Fig. 22(a) illustrates the tracking framework of BatTracker.

Extensive experiments are carried out in a highly cluttered lab to evaluate the performance of BatTracker. The results show that it can achieve device movement tracking in a 3D indoor space with the sub-cm accuracy and the 90-percentile error is less than 1 cm in a quiet scenario. Besides, the authors evaluate the performance in a 2D space by comparing BatTracker with CAT [78] and AAMouse [79]. The results of Fig. 22(b) show that BatTracker can achieve a maximum tracking error less than 1 cm and the tracking accuracy of BatTracker is higher than them. Moreover, the authors demonstrate that BatTracker can track any motion in a 3D space (see in Fig. 23). In the future work, BatTracker will concentrate on enhancing the tracking robustness from four processes: addressing the track loss issues, using effective information from other targets, utilizing customized microphones, and conducting more experiments on different types of mobiles.

6) STEERTRACK [131]

In 2018, X. Xu *et al.* present a device-free steering tracking system called SteerTrack [131]. This system leverages the ultrasonic signal of smartphone to track the hand trajectory

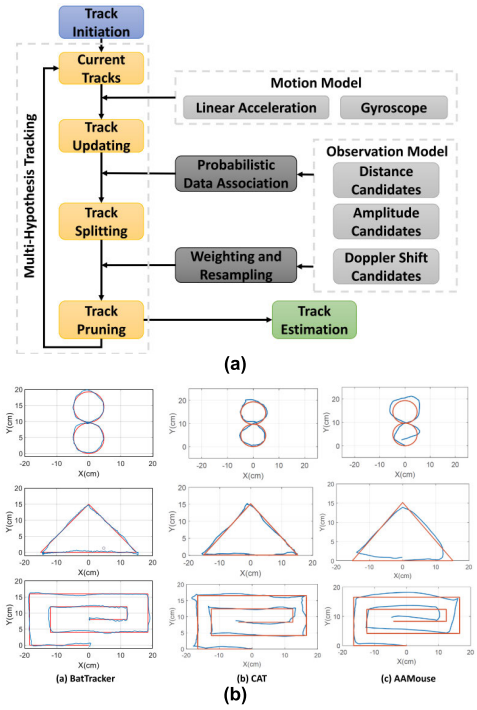


FIGURE 22. (a) The tracking framework of BatTracker; (b) The comparison results [132].

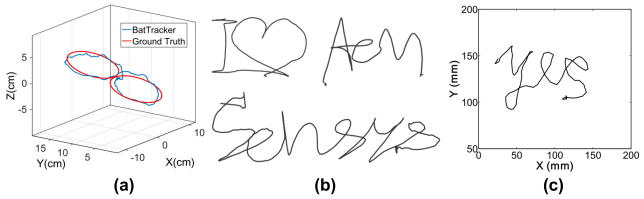


FIGURE 23. Tracking results in 3D space [132]. (a) Comparison result with ground truth; (b) Writing "I ♡ ACM Sensys"; (c) Drawing a spiral.

and estimate the rotation angle of steering wheel according to the tracked hand motions. It first utilizes the smartphone's embedded-in speaker to transmit a 20 kHz sinusoidal signal and two microphones built in smartphone to capture the echo signal reflected by hand motions. Then it uses relative correlation coefficient (RCC) and reference frame to analyze the echo signal to realize hand tracking. With the obtained hand movement trajectory, the authors propose an approach based on geometrical transformation that maps the steering wheel in 3D to 2D ellipse to estimate the rotation angle of the steering wheel. Fig. 24 shows the fundamental principle of tracking the rotation angle of steering wheel in 2D plane and the system framework of SteerTrack.

In the experimental phase, 5 different smartphones are deployed to evaluate SteerTrack with 5 participants in two real driving scenarios. The overall average absolute tracking error of SteerTrack is 4.61 degree and its median error is less than 4.79 for 5 drivers. It also assesses different steering maneuvers including Near hand, Farther hand, and both hands. For these three maneuvers, the recognition accuracy

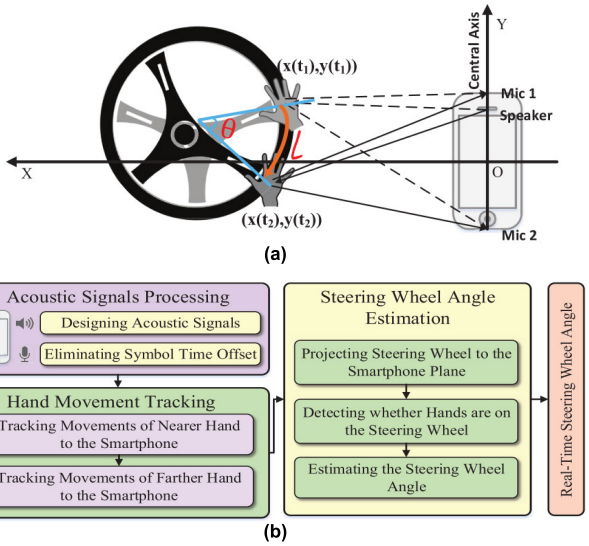


FIGURE 24. (a) Tracking the rotation angle of steering wheel by smartphone in 2D plane; (b) The system framework [131].

is 97.73%. Besides, it analyzes system performance under different road conditions (peak time and off-peak time) and road types (local road and highway). The absolute error is less than 10 degree under all four combination of road types and traffic conditions. The authors also verify that SteerTrack outperforms the other two methods, steering-wheel-mounted sensor [134] and smartwatch-based [135]. Furthermore, they study the influence of the smartphone position using 5 types of vehicles, including two different jeeps, two distinct cars, and van [136]. They place the smartphone on 5 positions (the left side, the middle side, and the right side of the instrument panel; near cab door; near cup-holder) in each vehicle, respectively. The experimental results demonstrate that SteerTrack performs well in all 5 vehicles when the smartphone is placed near the instrument panel. The future work of SteerTrack will further concentrate on improving the robustness of the placement of smartphone.

V. DISCUSSION

In this section, we compare the two types of hand gesture studies and emphasize the characteristics of their applications. We concentrate on the similarity and difference of these applications and present the analysis of the system performance from the following aspects, including signal types, feature presentation, experimental environments, and recognition accuracy. The findings facilitate the development of potential applications and provide some insights into this specific behavior recognition. The dynamic hand gesture recognition can be treated as a classification problem because we usually define a few specific hand gestures and classify the unknown gesture into these predefined labels. Differently, hand tracking usually comprises localization and tracking. The former locates the position of hand while the latter continuously determines the coordinate of hand in 2D or 3D space. Therefore, hand tracking is more challenging

because it must locate the position of hand more fined granularity to recognize some letters or shapes. We divide the discussion into three parts including signal acquisition, signal processing, and performance evaluation. After presenting the common features of dynamic gesture recognition and hand tracking, we illustrate the key characteristics of them and make a comprehensive analysis of these applications based on smartphone ultrasonic signals.

A. DYNAMIC GESTURE RECOGNITION

We investigate the gesture recognition and analyze various factors that affect recognition accuracy. We focus on the signal flow from ultrasonic signal collection to classification methods.

1) SIGNAL ACQUISITION

Android is the most popular smartphone operating system and has been adopted in numerous phone companies, such as MI, Samsung, Vivo, HTC, Sony, and Huawei. As a result, almost all dynamic hand gesture recognition systems choose Android smartphone as the experimental devices. Due to requirements of non-intrusive working pattern, we exploit the ultrasonic signal to recognize hand gestures. The frequency range varies from 16 kHz to 24 kHz, which can be excellently supported by the popular smartphone. The sound waveforms of these systems usually are continuous tones, such as sine wave, which owns some advantages such as simple frequency and low phase noise. Also, there are other types of modulation signals, which depends on the specific applications.

2) SIGNAL PROCESSING

After collecting ultrasound data, we need to eliminate noises from ambient factors and sensor devices and normalize the data. From Table 1, we find almost all dynamic hand gesture applications employ the Doppler effect. The possible reasons may be that this effect has been widely studied and has a satisfactory frequency resolution. Based on the principle, we can obtain the spectrum including time, frequency, and amplitude. The spectrum depicts the unique mapping relationship between the frequency variation and hand gestures. As a result, the specific gestures can be identified based on the unique rule. Next, the data are fed into a classifier to categorize the unknown hand gesture into a specific movement type. As shown in Table 2, the classification methods comprise common machine learning and deep learning algorithms, such as HMM, SVM, RNN, and CNN. These systems fully leverage the capability of general classification algorithms to determine hand gesture. These systems employ favorite classification algorithm based on the requirements of system design and extracted features.

3) PERFORMANCE EVALUATION

We evaluate the system performance from the following aspects including experimental environments, actions conducted, and recognition accuracy. From Table 2, to validate the algorithm performance, the experiments usually are carried out in various scenarios that keep different noise levels. For example, the levels of noise usually include quiet, normal, and noisy. And typical scenarios include home, lab, outdoor, restrictions, etc. Besides, the system performs many gestures to confirm the algorithm robustness. The least number of gestures of these systems is 5 and the largest number is up to 24, which proves that these algorithms have more adaptability. Besides, the number of participants varies from 3 to 11. The size of samples usually exceeds 900 and some size of systems reach 36000. As shown in Table 2, these systems achieve satisfactory accuracy, which is above 90%, and some of them up to 97%.

B. HAND TRAJECTORY TRACKING

Hand tracking is a challenging problem using ultrasonic signal because we have to locate the position of finger continuously under the weak signal changes. Dynamic gesture recognition systems usually leverage Doppler shift to extract movement features because these motions hold a large range of movements and can be easily distinguished. Different from these systems, hand tracking continuously requires more fine-grained localization and more low delay response. Therefore, how to extract weak sound signal from echo profile and how to convert these signals into hand movement trajectory are the major challenges. Each system employs an appropriate solution to address these problems. We investigate the state-of-the-art applications and present the specific solution to these difficulties. We reveal the crucial idea and analyze the characteristics of hand tracking applications from the following aspects: signal acquisition, signal processing, and performance evaluation.

1) SIGNAL ACQUISITION

The hardware devices employed in these systems are similar to that of dynamic gesture recognition. The experimental devices of the tracking system usually apply Android system from various companies. As we have analyzed, simple sinusoidal sound signal cannot provide enough resolution to the hand tracking as the range of hand or finger may be very short and the change of signal is too weak to be measured effectively. Therefore, these systems usually employ special sound signal modulation such as OFDM, BPSK, chirp, etc. because these modulations provide more favorite features. Besides, some systems leverage the two speakers or microphones to track the finger in 2D or 3D space. Different from dynamic gesture recognition that usually applies Doppler shift, these tracking systems exploit many different features including phase, ToF, CIR, and Doppler shift.

2) SIGNAL PROCESSING

Although different signal modulation methods are employed, the means of signal preprocessing is similar. These preprocessing methods include band-pass filter, noise elimination, and data normalization. Besides, some specific approaches are applied according to the signal types, such as I/Q demodulation for FingerIO [111] and multipath effect elimination for

EchoTrack [130]. The hand tracking systems usually leverage geometric models to achieve the last result rather than common classifiers because it needs to continuously calculate fine-grained location to identify some characters, words, or shapes. Although some general methods such as particle filter are widely used in various localization scenarios, they usually are not employed in hand tracking scenarios due to its computation cost for smartphone.

3) PERFORMANCE EVALUATION

We assess the performance of these systems from localization in 1D space and tracking in 2D and 3D space, which is a noticeable difference from dynamic gesture recognition. Besides the common experimental environments, we consider the additional criteria including time delay, power consumption, and shape or character recognition in 2D and 3D space. We evaluate the performance of these tracking systems from these aspects.

The most systems apply laboratory as the test environment and some systems use offices or outdoor with some noises to validate the robustness of algorithms. Some systems apply specific environment based on their application scenarios. For example, SteerTrack [131] employs real driving environment on local road and highway. From the number of participants, the most are 20 users from EchoTrack [130]. Most of them have 5 to 10 users. From tracking behavior, different applications validate performance using different hand behaviors. For example, LLAP [128] uses 26 letters and 11 words, which is the most complex hand recognition among these studies. FingerIO [111] recognizes all kinds of shapes and Strata identifies three types of shapes. Different from them, StreeTrack [131] tracks the angle of steer rotating. We find it is difficult to compare these algorithms because they perform various behaviors under different environments.

From the analysis of Section IV, we find that hand tracking involves many aspects and they affect recognition accuracy. Firstly, different evaluation metrics usually are employed because these applications need to estimate algorithm performance. For example, LLAP [128] estimates system performance from many aspects from 1D and 2D spaces. For 1D space, the assessment criteria of LLAP [128] includes movement distance error, absolute path length error, and movement detection precision. For 2D space, it uses tracking error and character recognition accuracy as metrics. Differently, FingerIO [111] applies cumulative distribution functions (CDFs) of 2D tracking errors as evaluation metrics. These metrics illustrate system performance based on the characteristics of applications. Secondly, it is challenging to compare different algorithm performance because many systems have specific functions. For example, although Strata [129] compares the distance error and trajectory error with LLAP [128] and FingerIO [111], the letter recognition in LLAP [128] and occluded scenario in FingerIO [111] are not provided because they are the pivotal features different from other systems. Thirdly, some systems track hand movement in 3D space while other systems in 2D. Therefore, the former has to face more difficulties due to more dimension of the movement, which leads to the decrease of tracking precision. Lastly, although many systems evaluate recognition accuracy using noise environment, the source and the level of noise are different. Therefore, the performance comparison in these scenarios just confirms the effectiveness of the algorithm under their environment, which may not be suitable for other scenarios or studies.

VI. CHALLENGES

Currently, smartphone has become the most popular electronic device for most people because it provides us with various functions and enriches our lives. Besides common related sound functions, the built-in speakers and microphones can emit and receive ultrasonic signal, which can make smartphone serve as an active sonar sensing system. Recently, many interesting applications of hand gesture recognition based on this sensing system has been developed and deployed in many scenarios. Although these systems achieve good recognition performance, there are many issues to be addressed to improve identification accuracy. In this section, we will present some main issues and possible solutions to the active sonar hand gesture sensing applications based on smartphone.

A. MULTI-DATA FUSION

Most applications in this paper adopt one or two microphones and speakers built in the smartphone to recognize hand gesture. The number of sensors limits the size of the measurement data and constrains the types of gestures. Therefore, these systems recognize a few types of dynamic gestures and track hand in 1D and 2D space. UltraGesture [118] proves that adding extra sensors can achieve fine-grained gesture recognition and identify more hand postures. To recognize more hand postures and track hand in a 3D space, using more speakers and microphones would be a fine solution [131]. Besides, the fusion of many features, including Doppler shift, phase, ToF, and CIR would be another simple and effective method. Other sensors embedded in the smartphone (e.g., accelerometer [137], gyroscope) can also provide more useful information to improve the quality of HCI and recognition accuracy.

B. ROBUSTNESS

Current studies validate system performance under various environments which have different noise levels. However, it seems difficult to compare these levels of noise because they are much different. Besides, human movements near the participant severely affect recognition accuracy because the movements generate complicated multipath effect [111]. In addition, the distance between the smartphone and the user is also an important factor that affects the recognition performance [128]. Moreover, the hand size, the hand that participants used (left or right), the hand with or without occlusion can also impact the accuracy of the application [118]. Therefore, how to eliminate the impacts of

nearby environmental changes and interference is a challenging problem.

C. STANDARD DATASET

Currently, many systems evaluate the performance by using different types of actions. For dynamic gesture recognition, the number of gestures is very different [112], [113]. It varies from 5 to 24. For hand tracking, the shapes drawn by hand contain some simple shapes (e.g., diamond, triangle, circle) or letters and words [128], [129]. The variation of the shapes generates significant difficulty when evaluating the system performance because of the difference of the experimental conditions. To effectively compare the performance of various algorithms, the standard dataset is required.

D. LOW-LATENCY

For gesture recognition applications, latency usually is not a crucial metric because we pay more attention to recognition accuracy. However, for hand tracking applications, we may concentrate on the latency since many applications require a swift response to hand position. For example, low-latency is a crucial factor for real-time game control. However, a few applications consider latency feature when evaluating system performance, such as Strata [129] and LLAP [128]. Most others do not analyze the latency factor when testing system performance. Besides that, the signal types and feature calculation are must be considered when developing a hand tracking algorithm.

E. SECURITY ISSUES

Hand gesture recognition based on the ultrasonic signal of smartphone brings us a novel method for HCI. Since the systems do not require any extra hardware devices and provide us with convenient interaction means for users, researchers pay more attention to the development of applications based on ultrasonic signal. However, this method is also accompanied by security issues. Because the hearing range of the average people is within 16 kHz [133] and our presented systems adopt the sound signal with frequency ranging 16 kHz - 24 kHz, a potential attack using ultrasonic signal can be launched. At the same time, this method might be utilized to steal people's information, infer Android unlock patterns [115], [124], even carry out inaudible sound attack [39]. Therefore, some practical approaches need to be studied in the future to tackle these issues.

VII. CONCLUSION

With the rapid development of hardware and software technology, smartphone is becoming a powerful communication and entertainment tool and has become an indispensable electric device in daily lives because it provides us with various functions to facilitate our work, lives, entertainment, and mutual communications. Nowadays, the speakers and microphones built in smartphone have better performance than before; therefore, many researchers explore the active sonar technique using the smartphone's built-in speakers and microphones to recognize hand gestures. This active sonar sensing system uses smartphone to emit and capture ultrasonic signal and analyzes the echo signal to realize dynamic hand gesture recognition and hand tracking. Since the ultrasonic signal would not interfere with people's normal life, using smartphone-based ultrasonic signal to classify hand gestures can facilitate the development of the general and long-term tracking system. Besides, this system provides us with a convenient and natural HCI method and enriches the IoT applications.

The purpose of this paper is to present a comprehensive survey on the state-of-the-art applications of hand gestures based on active sonar sensing system using smartphone. This paper concentrates on the crucial characteristics of the framework of the ultrasonic sensing system and analyzes the related applications from dynamic gesture recognition and hand trajectory tracking. Firstly, we review some common hand gesture recognition systems based on acoustic signal and categorize the smartphone-based audio sensing systems into two groups including passive acoustic sensing and active acoustic sensing. Secondly, we propose a typical framework based on the hand recognition applications by using ultrasonic signal and analyze the basic principles of the hand gesture recognition system based on ultrasonic signal of smartphone. Thirdly, we introduce some processing techniques adopted in hand gesture recognition, including signal generation, signal extraction, noise elimination, and hand gesture recognition methods. Next, this paper investigates the state-of-the-art applications about dynamic hand gesture recognition and hand tracking. For each application, we review the key features and analyze the system performance. Then we make a detailed discussion about these systems from signal acquisition, signal processing, and performance evaluation. Finally, based on current study trends, we discuss the limitations and open issues involved in human hand gesture recognition based on the ultrasonic signal of smartphone and present some potential solutions to these issues.

REFERENCES

[1] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, pp. 1005:1–1005:20, 2019.

[2] C. Cai, R. Zheng, and M. Hu, "A survey on acoustic sensing," 2019, *arXiv:1901.03450*. [Online]. Available: https://arxiv.org/abs/1901.03450

[3] Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with channel state information: A survey," *ACM Comput. Surveys*, vol. 52, no. 3, Jun. 2019, Art. no. 46.

[4] R. H. Venkatnarayan and M. Shahzad, "Gesture recognition using ambient light," *ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, Mar. 2018, Art. no. 40.

[5] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.

[6] B. Fu, F. Kirchbuchner, A. Kuijper, A. Braun, and D. V. Gangatharan, "Fitness activity recognition on smartphones using Doppler measurements," *Informatics*, vol. 5, no. 2, p. 24, May 2018.

[7] B. Fu, D. V. Gangatharan, A. Kuijper, F. Kirchbuchner, and A. Braun, "Exercise monitoring on consumer smart phones using ultrasonic sensing," in *Proc. 4th Int. Workshop Sensor-based Activity Recognit. Interact.*, Rostock, Germany, 2017, Art. no. 9.

[8] X. Guo, X. Hu, X. Ye, C. Hu, C. Song, and H. Wu, "Human activity recognition based on two-dimensional acoustic arrays," in *Proc. IEEE Int. Ultrasonics Symp. (IUS)*, Kobe, Japan, Oct. 2018, pp. 1–4.

[9] A. Ghosh, A. Chakraborty, D. Chakraborty, M. Saha, and S. Saha, "UltraSense: A non-intrusive approach for human activity identification using heterogeneous ultrasonic sensor grid for smart home environment," *J. Ambient Intell. Hum. Comput.*, pp. 1–22, Mar. 2019.

[10] A. Ghosh, D. Chakraborty, D. Prasad, M. Saha, and S. Saha, "Can we recognize multiple human group activities using ultrasonic sensors?" in *Proc. 10th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Bengaluru, India, Jan. 2018, pp. 557–560.

[11] R. Nandakumar, A. Takakuwa, T. Kohno, and S. Gollakota, "CovertBand: Activity information leakage using music," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, Sep. 2017, Art. no. 87.

[12] T. S. Murray, D. R. Mendat, K. A. Sanni, P. O. Pouliquen, and A. G. Andreou, "Bio-inspired human action recognition with a micro-Doppler sonar system," *IEEE Access*, vol. 6, pp. 28388–28403, 2017.

[13] A. Agarwal, M. Jain, P. Kumar, and S. Patel, "Opportunistic sensing with MIC arrays on smart speakers for distal interaction and exercise tracking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 6403–6407.

[14] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "StressSense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proc. ACM Conf. Ubiquitous Comput.*, Pittsburgh, PA, USA, 2012, pp. 351–360.

[15] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, "Accurate and privacy preserving cough sensing using a low-cost microphone," in *Proc. 13th Int. Conf. Ubiquitous Comput.*, Beijing, China, 2011, pp. 375–384.

[16] T. Wang, D. Zhang, L. Wang, Y. Zheng, T. Gu, B. Dorizzi, and X. Zhou, "Contactless respiration monitoring using ultrasound signal with off-the-shelf audio devices," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2959–2973, Apr. 2019.

[17] K. Qian, C. Wu, F. Xiao, Y. Zheng, Y. Zhang, Z. Yang, and Y. Liu, "Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1574–1582.

[18] R. Nandakumar, S. Gollakota, and J. E. Sunshine, "Opioid overdose detection using smartphones," *Sci. Transl. Med.*, vol. 11, no. 474, 2019, Art. no. eaau8914.

[19] J. Chan, T. Rea, S. Gollakota, and J. E. Sunshine, "Contactless cardiac arrest detection using smart devices," 2019, *arXiv:1902.00062v2*. [Online]. Available: https://arxiv.org/abs/1902.00062v2

[20] J. Chan, S. Raju, R. Nandakumar, R. Bly, and S. Gollakota, "Detecting middle ear fluid using smartphones," *Sci. Transl. Med.*, vol. 11, no. 492, 2019, Art. no. eaav1102.

[21] R. Huang, "Contact-free breathing rate monitoring with smartphones: A sonar phase approach," M.S. thesis, Auburn Univ., Auburn, AL, USA, Aug. 2018.

[22] L. Ge, J. Zhang, and J. Wei, "Single-frequency ultrasound-based respiration rate estimation with smartphones," *Comput. Math. Methods Med.*, vol. 2018, pp. 1–8, May 2018.

[23] R. Nandakumar and S. Gollakota, "Unleashing the power of active sonar," *IEEE Pervasive Comput.*, vol. 16, no. 1, pp. 11–15, Jan./Mar. 2017.

[24] B. Liu, X. Dai, H. Gong, Z. Guo, N. Liu, X. Wang, and M. Liu, "Deep learning versus professional healthcare equipment: A fine-grained breathing rate monitoring model," *Mobile Inf. Syst.*, vol. 2018, pp. 1–9, Mar. 2018.

[25] Y. Yamada, K. Shinkawa, T. Takase, A. Kosugi, K. Fukuda, and M. Kobayashi, "Monitoring daily physical conditions of older adults using acoustic features: A preliminary result," *Stud. Health Technol. Inf.*, vol. 247, pp. 301–305, May 2018.

[26] T. Wang, D. Zhang, Y. Zheng, T. Gu, X. Zhou, and B. Dorizzi, "C-FMCW based contactless respiration detection using acoustic signal," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, Dec. 2017, Art. no. 170.

[27] S. D. Min, J. K. Kim, H. S. Shin, Y. H. Yun, C. K. Lee, and M. Lee, "Noncontact respiration rate measurement system using an ultrasonic proximity sensor," *IEEE Sensors J.*, vol. 10, no. 11, pp. 1732–1739, Nov. 2010.

[28] X. Wang, R. Huang, and S. Mao, "SonarBeat: Sonar phase for breathing beat monitoring with smartphones," in *Proc. 26th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Vancouver, BC, Canada, Jul./Aug. 2017, pp. 1–8.

[29] Y. Ren, C. Wang, J. Yang, and Y. Chen, "Fine-grained sleep monitoring: Hearing your breathing with smartphones," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Hong Kong, Apr./May 2015, pp. 1194–1202.

[30] J. Tan, C.-T. Nguyen, and X. Wang, "SilentTalk: Lip reading through ultrasonic sensing on mobile phones," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun.*, Atlanta, GA, USA, May 2017, pp. 1–9.

[31] J. Tan, X. Wang, C.-T. Nguyen, and Y. Shi, "SilentKey: A new authentication framework through ultrasonic-based lip reading," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, Mar. 2018, Art. no. 36.

[32] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1466–1474.

[33] Y. Zou, M. Zhao, Z. Zhou, J. Lin, M. Li, and K. Wu, "BiLock: User authentication via dental occlusion biometrics," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, Sep. 2018, Art. no. 152.

[34] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "EchoPrint: Two-factor authentication using acoustics and vision on smartphones," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, New Delhi, India, 2018, pp. 321–336.

[35] H. Muckenhirn, M. Magimai-Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using CNNS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 4884–4888.

[36] Y. Wang, Y. Chen, M. Z. A. Bhuiyan, Y. Han, S. Zhao, and J. Li, "Gait-based human identification using acoustic sensor and deep neural network," *Future Gener. Comput. Syst.*, vol. 86, pp. 1228–1237, Sep. 2018.

[37] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "BreathPrint: Breathing acoustics-based user authentication," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, Niagara Falls, NY, USA, 2017, pp. 278–291.

[38] N. Khalil, D. Benhaddou, O. Gnawali, and J. Subhlok, "Nonintrusive ultrasonic-based occupant identification for energy efficient smart building applications," *Appl. Energy*, vol. 220, pp. 814–828, Mar. 2018.

[39] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, "Inaudible voice commands: The long-range attack and defense," in *Proc. 15th USENIX Conf. Netw. Syst. Design Implement.*, Renton, WA, USA, 2018, pp. 547–560.

[40] I. Shumailov, L. Simon, J. Yan, and R. Anderson, "Hearing your touch: A new acoustic side channel on smartphones," 2019, *arXiv:1903.11137*. [Online]. Available: https://arxiv.org/abs/1903.11137

[41] C. Gao, K. Fawaz, S. Sur, and S. Banerjee, "Privacy protection for audio sensing against multi-microphone adversaries," *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 2, pp. 146–165, Dec. 2019.

[42] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, Snowbird, UT, USA, 2017, pp. 343–355.

[43] T. Yu, H. Jin, and K. Nahrstedt, "WritingHacker: Audio based eavesdropping of handwriting via mobile devices," in *Proc. 2016 ACM Int. Joint Conf. Pervas. Ubiquitous Comput.*, Berlin, Germany, 2016, pp. 463–473.

[44] K. Sun, W. Wang, A. X. Liu, and H. Dai, "Depth aware finger tapping on virtual displays," in *Proc. 16th Annu. Int. Conf. Mobile Syst., Appl., Services*, Munich, Germany, 2018, pp. 283–295.

[45] C. Peng, G. Shen, and Y. Zhang, "BeepBeep: A high-accuracy acoustic-based system for ranging and localization using COTS devices," *ACM Trans. Embedded Comput. Syst.*, vol. 11, no. 1, Mar. 2012, Art. no. 4.

[46] Z. Zhang, D. Chu, X. Chen, and T. Moscibroda, "SwordFight: Enabling a new class of phone-to-phone action games on commodity phones," in *Proc. 10th Int. Conf. Mobile Syst., Appl., Services*, Low Wood Bay, U.K., 2012, pp. 1–14.

[47] W. Huang, Y. Xiong, X.-Y. Li, H. Lin, X. Mao, P. Yang, Y. Liu, and X. Wang, "Swadloon: Direction finding and indoor localization using acoustic signal by shaking smartphones," *IEEE Trans. Mobile Comput.*, vol. 14, no. 10, pp. 2145–2157, Oct. 2015.

[48] Y.-C. Tung and K. G. Shin, "EchoTag: Accurate infrastructure-free indoor location tagging with smartphones," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Paris, France, 2015, pp. 525–536.

[49] H. Chen, F. Li, and Y. Wang, "EchoLoc: Accurate device-free hand localization using COTS devices," in *Proc. 45th Int. Conf. Parallel Process. (ICPP)*, Philadelphia, PA, USA, Aug. 2016, pp. 334–339.

[50] K. Liu, X. Liu, and X. Li, "Guoguo: Enabling fine-grained smartphone localization via acoustic anchors," *IEEE Trans. Mobile Comput.*, vol. 15, no. 5, pp. 1144–1156, May 2016.
[51] M. A. Zayer, S. Tregillus, and E. Folmer, "PAWdio: Hand input for mobile VR using acoustic sensing," in *Proc. 2016 Annu. Symp. Comput.-Hum. Interact. Play*, Austin, TX, USA, 2016, pp. 154–158.
[52] C. Liu, S. Jiang, S. Zhao, and Z. Guo, "Infrastructure-free indoor pedestrian tracking with smartphone acoustic-based enhancement," *Sensors*, vol. 19, no. 11, p. 2458, May 2019.
[53] Q. Lin, Z. An, and L. Yang, "Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices," in *Proc. MobiCom*, Los Cabos, Mexico, 2019, pp. 1–16.
[54] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "BatMapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, Niagara Falls, NY, USA, 2017, pp. 42–55.
[55] S. Pradhan, G. Baig, W. Mao, L. Qiu, G. Chen, and B. Yang, "Smartphone-based acoustic indoor space mapping," *ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 2, Jun. 2018, Art. no. 75.
[56] W. Mao, M. Wang, and L. Qiu, "AIM: Acoustic imaging on a mobile," in *Proc. 16th Annu. Int. Conf. Mobile Syst., Appl., Services*, Munich, Germany, 2018, pp. 468–481.
[57] F. Li, H. Chen, X. Song, Q. Zhang, Y. Li, and Y. Wang, "CondioSense: High-quality context-aware service for audio sensing system via active sonar," *Pers. Ubiquitous Comput.*, vol. 21, no. 1, pp. 17–29, Feb. 2017.
[58] Q. Song, C. Gu, and R. Tan, "Deep room recognition using inaudible echos," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 3, Sep. 2018, Art. no. 135.
[59] I. Bisio, A. Delfino, A. Grattarola, F. Lavagetto, and A. Sciarrone, "Ultrasounds-based context sensing method and applications over the Internet of Things," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3876–3890, Oct. 2018.
[60] M. Rossi, J. Seiter, O. Amft, S. Buchmeier, G. Tröster, "RoomSense: An indoor positioning system for smartphones using active sound probing," in *Proc. 4th Augmented Hum. Int. Conf.*, Stuttgart, Germany, 2013, pp. 89–95.
[61] M. Ono, B. Shizuki, and J. Tanaka, "Sensing touch force using active acoustic sensing," in *Proc. 9th Int. Conf. Tangible, Embedded, Embodied Interact.*, Stanford, CA, USA, 2015, pp. 355–358.
[62] F. Alonso-Martín, J. J. Gamboa-Montero, J. C. Castillo, Á. Castro-González, and M. Á. Salichs, "Detecting and classifying human touches in a social robot through acoustic sensing and machine learning," *Sensors*, vol. 17, no. 5, p. 1138, May 2017.
[63] E. W. Pedersen and K. Hornbæk, "Expressive touch: Studying tapping force on tabletops," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Toronto, ON, Canada, 2014, pp. 421–430.
[64] Z. Li, H. Dai, W. Wang, A. X. Liu, and G. Chen, "PCIAS: Precise and contactless measurement of instantaneous angular speed using a smartphone," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 4, Dec. 2018, Art. no. 177.
[65] Z. Sun, R. Bose, and P. Zhang, "Spartacus: Spatially-aware interaction for mobile devices through energy-efficient audio sensing," *GetMobile, Mobile Comput. Commun.*, vol. 18, no. 4, pp. 11–14, Oct. 2014.
[66] M. T. I. Aumi, S. Gupta, M. Goel, E. Larson, and S. Patel, "DopLink: Using the Doppler effect for multi-device interaction," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Zürich, Switzerland, 2013, pp. 583–586.
[67] P. Xie, J. Feng, Z. Cao, and J. Wang, "GeneWave: Fast authentication and key agreement on commodity mobile devices," *IEEE/ACM Trans. Netw.*, vol. 26, no. 4, pp. 1688–1700, Aug. 2018.
[68] H. Jin, C. Holz, and K. Hornbæk, "Tracko: Ad-hoc mobile 3D tracking using Bluetooth low energy and inaudible signals for cross-device interaction," in *Proc. 28th Annu. ACM Symp. User Interface Softw. Technol.*, Charlotte, NC, USA, 2015, pp. 147–156.
[69] G. Shakeri, J. H. Williamson, and S. Brewster, "May the force be with you: Ultrasound haptic feedback for mid-air gesture interaction in cars," in *Proc. 10th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, Toronto, ON, Canada, 2018, pp. 1–10.
[70] X. Xu, J. Yu, Y. Chen, Y. Zhu, S. Qian, and M. Li, "Leveraging audio signals for early recognition of inattentive driving with smartphones," *IEEE Trans. Mobile Comput.*, vol. 17, no. 7, pp. 1553–1567, Jul. 2018.
[71] Z. Wang, S. Tan, L. Zhang, and J. Yang, "ObstacleWatch: Acoustic-based obstacle collision detection for pedestrian using smartphone," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 4, Dec. 2018, Art. no. 194.
[72] Y.-C. Tung and K. G. Shin, "Use of phone sensors to enhance distracted pedestrians' safety," *IEEE Trans. Mobile Comput.*, vol. 17, no. 6, pp. 1469–1482, Jun. 2018.
[73] Q. Zhang, J. Lin, H. Song, and G. Sheng, "Fault identification based on PD ultrasonic signal using RNN, DNN and CNN," in *Proc. Condition Monit. Diagnosis (CMD)*, Perth, WA, Australia, Sep. 2018, pp. 1–6.
[74] P. Rzeszucinski, M. Orman, C. T. Pinto, A. Tkaczyk, and M. Sulowicz, "Bearing health diagnosed with a mobile phone: Acoustic signal measurements can be used to test for structural faults in motors," *IEEE Ind. Appl. Mag.*, vol. 24, no. 4, pp. 17–23, Jul./Aug. 2018.
[75] R. Kapoor, S. Ramasamy, A. Gardi, R. Van Schyndel, and R. Sabatini, "Acoustic sensors for air and surface navigation applications," *Sensors*, vol. 18, no. 2, p. 499, Feb. 2018.
[76] A. Wang and S. Gollakota, "MilliSonic: Pushing the limits of acoustic motion tracking," in *Proc. CHI*, Glasgow, U.K., 2019, pp. 1–11.
[77] D. Ren, Y. Zhang, N. Xiao, H. Zhou, X. Li, J. Qian, and P. Yang, "WordFi: Accurate handwrite system empowered by wireless backscattering and machine learning," *IEEE Netw.*, vol. 32, no. 4, pp. 47–53, Jul./Aug. 2018.
[78] W. Mao, J. He, and L. Qiu, "CAT: High-precision acoustic motion tracking," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, New York, NY, USA, 2016, pp. 69–81.
[79] S. Yun, Y.-C. Chen, and L. Qiu, "Turning a mobile device into a mouse in the air," in *Proc. 13th Annu. Int. Conf. Mobile Syst., Appl., Services*, Florence, Italy, 2015, pp. 15–29.
[80] W. Liu, W. Shen, B. Li, and L. Wang, "Toward device-free micro-gesture tracking via accurate acoustic Doppler-shift detection," *IEEE Access*, vol. 7, pp. 1084–1094, 2018.
[81] M. Zhang, Q. Dai, P. Yang, J. Xiong, C. Tian, and C. Xiang, "iDial: Enabling a virtual dial plate on the hand back for around-device interaction," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 1, Mar. 2018, Art. no. 55.
[82] M. Chen, P. Yang, S. Cao, M. Zhang, and P. Li, "WritePad: Consecutive number writing on your hand with smart acoustic sensing," *IEEE Access*, vol. 6, pp. 77240–77249, 2018.
[83] S. Cao, P. Yang, X. Li, M. Chen, and P. Zhu, "iPand: Accurate gesture input with smart acoustic sensing on hand," in *Proc. 15th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Hong Kong, Jun. 2018, pp. 1–3.
[84] C. R. Pittman and J. J. LaViola, Jr., "Multiwave: Complex hand gesture recognition using the Doppler effect," in *Proc. 43rd Graph. Interface Conf.*, Edmonton, AB, Canada, 2017, pp. 97–106.
[85] C. Pittman, P. Wisniewski, C. Brooks, and J. J. LaViola, Jr., "Multiwave: Doppler effect based gesture recognition in multiple dimensions," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, San Jose, CA, USA, 2016, pp. 1729–1736.
[86] Q. Liu, W. Yang, Y. Xu, Y. Hu, Q. He, and L. Huang, "DopGest: Dual-frequency based ultrasonic gesture recognition," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Guangzhou, China, Oct. 2018, pp. 293–300.
[87] S. Gupta, D. Morris, S. Patel, and D. Tan, "SoundWave: Using the Doppler effect to sense gestures," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Austin, TX, USA, 2012, pp. 1911–1914.
[88] L. Lu, J. Liu, J. Yu, Y. Chen, Y. Zhu, X. Xu, and M. Li, "VPad: Virtual writing tablet for laptops leveraging acoustic signals," in *Proc. IEEE 24th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Singapore, Dec. 2018, pp. 244–251.
[89] H. Ai, Y. Men, L. Han, Z. Li, and M. Liu, "High precision gesture sensing via quantitative characterization of the Doppler effect," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancún, Mexico, Dec. 2016, pp. 973–978.
[90] H. Ai, L. Han, Y. Wang, and L. Liao, "Accurate acoustic based gesture classification with zero start-up cost," in *Proc. ICA3PP*, Guangzhou, China, 2018, pp. 44–58.
[91] C. Zhang, Q. Xue, A. Waghmare, R. Meng, S. Jain, Y. Han, X. Li, K. Cunefare, T. Ploetz, T. Starner, O. Inan, and G. D. Abowd, "FingerPing: Recognizing fine-grained hand poses using active acoustic on-body sensing," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Montreal QC, Canada, 2018, pp. 1–10.

[92] Y. Iravantchi, M. Goel, and C. Harrison, "BeamBand: Hand gesture sensing with ultrasonic beamforming," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Glasgow, U.K., 2019, pp. 1–10.

[93] C. Zhang, Q. Xue, A. Waghmare, S. Jain, Y. Pu, S. Hersek, K. Lyons, K. A. Cunefare, O. T. Inan, and G. D. Abowd, "SoundTrak: Continuous 3D tracking of a finger using active acoustics," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 2, Jun. 2017, Art. no. 30.

[94] H. Chen, T. Ballal, M. Saad, and T. Y. Al-Naffouri, "Angle-of-arrival-based gesture recognition using ultrasonic multi-frequency signals," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Kos, Greece, Aug./Sep. 2017, pp. 16–20.

[95] R. Xiao, G. Lew, J. Marsanico, D. Hariharan, S. Hudson, and C. Harrison, "Toffee: Enabling ad hoc, around-device interaction with acoustic time-of-arrival correlation," in *Proc. 16th Int. Conf. Hum.-Comput. Interact. Mobile Devices, Services*, Toronto, ON, Canada, 2014, pp. 67–76.

[96] Y. Sang, L. Shi, and Y. Liu, "Micro hand gesture recognition system using ultrasonic active sensing," *IEEE Access*, vol. 6, pp. 49339–49347, 2018.

[97] K. Kalgaonkar and B. Raj, "One-handed gesture recognition using ultrasonic Doppler sonar," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 1889–1892.

[98] Q. Zeng, Z. Kuang, S. Wu, and J. Yang, "A method of ultrasonic finger gesture recognition based on the micro-Doppler effect," *Appl. Sci.*, vol. 9, no. 11, p. 2314, Jun. 2019.

[99] J. Wang, R. Ruby, L. Wang, and K. Wu, "Accurate combined keystrokes detection using acoustic signals," in *Proc. 12th Int. Conf. Mobile Ad-Hoc Sensor Netw. (MSN)*, Hefei, China, Dec. 2016, pp. 9–14.

[100] J. Wang, K. Zhao, X. Zhang, and C. Peng, "Ubiquitous keyboard for small mobile devices: Harnessing multipath fading for fine-grained keystroke localization," in *Proc. 12th Annu. Int. Conf. Mobile Syst., Appl., Services*, Bretton Woods, NH, USA, 2014, pp. 14–27.

[101] J. Liu, Y. Wang, G. Kar, Y. Chen, J. Yang, and M. Gruteser, "Snooping keystrokes with mm-level Audio ranging on a single phone," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Paris, France, 2015, pp. 142–154.

[102] M. Zhang, P. Yang, C. Tian, L. Shi, S. Tang, and F. Xiao, "SoundWrite: Text input on surfaces through mobile acoustic sensing," in *Proc. 1st Int. Workshop Exper. Design Implement. Smart Objects*, Paris, France, 2015, pp. 13–17.

[103] G. Luo, M. Chen, P. Li, M. Zhang, and P. Yang, "SoundWrite II: Ambient acoustic sensing for noise tolerant device-free gesture recognition," in *Proc. IEEE 23rd Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Shenzhen, China, Dec. 2017, pp. 121–126.

[104] H. Yin, A. Zhou, L. Liu, N. Wang, and H. Ma, "Ubiquitous writer: Robust text input for small mobile devices via acoustic sensing," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5285–5296, Jun. 2019.

[105] M. Chen, P. Yang, J. Xiong, M. Zhang, Y. Lee, C. Xiang, and C. Tian, "Your table can be an input panel: Acoustic-based device-free interaction recognition," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 1, Mar. 2019, Art. no. 3.

[106] Z. Yu, H. Du, D. Xiao, Z. Wang, Q. Han, and B. Guo, "Recognition of human computer operations based on keystroke sensing by smartphone microphone," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1156–1168, Apr. 2018.

[107] Y. Zhang, J. Wang, W. Wang, Z. Wang, and Y. Liu, "Vernier: Accurate and fast acoustic motion tracking using mobile devices," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1709–1717.

[108] H. Du, P. Li, H. Zhou, W. Gong, G. Luo, and P. Yang, "WordRecorder: Accurate acoustic-based handwriting recognition using deep learning," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1448–1456.

[109] T. Zhu, Q. Ma, S. Zhang, and Y. Liu, "Context-free attacks using keyboard acoustic emanations," in *Proc. 2014 ACM SIGSAC Conf. Comput. Commun. Secur.*, Scottsdale, AZ, USA, 2014, pp. 453–464.

[110] H. Du, Z. Yu, D. Xiao, Z. Wang, Q. Han, and B. Guo, "Sensing keyboard input for computer activity recognition with a smartphone," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. ACM Int. Symp. Wearable Comput.*, Maui, HI, USA, 2017, pp. 25–28.

[111] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO: Using active sonar for fine-grained finger tracking," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Santa Clara, CA, USA, 2016, pp. 1515–1525.

[112] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan, "AudioGest: Enabling fine-grained hand gesture detection by decoding echo signal," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Berlin, Germany, 2016, pp. 474–485.

[113] Y. Qifan, T. Hao, Z. Xuebing, L. Yin, and Z. Sanfeng, "Dolphin: Ultrasonic-based gesture recognition on smartphone platform," in *Proc. IEEE 17th Int. Conf. Comput. Sci. Eng.*, Chengdu, China, Dec. 2014, pp. 1461–1468.

[114] Q. Yang, H. Fu, Y. Zou, and K. Wu, "A novel finger-assisted touch-free text input system without training," in *Proc. 16th Annu. Int. Conf. Mobile Syst., Appl., Services*, Munich, Germany, 2018, p. 533.

[115] P. Cheng, I. E. Bagci, U. Roedig, and J. Yan, "SonarSnoop: Active acoustic side-channel attacks," 2018, *arXiv:1808.10250*. [Online]. Available: https://arxiv.org/abs/1808.10250

[116] H. Watanabe and T. Terada, "Improving ultrasound-based gesture recognition using a partially shielded single microphone," in *Proc. ACM Int. Symp. Wearable Comput.*, Singapore, 2018, pp. 9–16.

[117] X. Li, H. Dai, L. Cui, and Y. Wang, "SonicOperator: Ultrasonic gesture recognition with deep neural network on mobiles," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, San Francisco, CA, USA, Aug. 2017, pp. 1–7.

[118] K. Ling, H. Dai, Y. Liu, and A. X. Liu, "UltraGesture: Fine-grained gesture sensing and recognition," in *Proc. 15th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Hong Kong, Jun. 2018, pp. 1–9.

[119] K.-Y. Chen, D. Ashbrook, M. Goel, S.-H. Lee, and S. Patel, "AirLink: Sharing files between multiple devices using in-air gestures," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Seattle, WA, USA, 2014, pp. 565–569.

[120] K. Sun, T. Zhao, W. Wang, and L. Xie, "VSkin: Sensing touch gestures on surfaces of mobile devices using acoustic signals," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, New Delhi, India, 2018, pp. 591–605.

[121] Y.-C. Tung and K. G. Shin, "Expansion of human-phone interface by sensing structure-borne sound propagation," in *Proc. 14th Annu. Int. Conf. Mobile Syst., Appl., Services*, Singapore, 2016, pp. 277–289.

[122] N. Kim and J. Lee, "Towards grip sensing for commodity smartphones through acoustic signature," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. ACM Int. Symp. Wearable Comput.*, Maui, HI, USA, 2017, pp. 105–108.

[123] Z. Xu, K. Wu, and P. Hu, "Ultrasonic waves based gesture recognition method for smartphone platform," *Comput. Eng. Appl.*, vol. 54, no. 2, pp. 239–245, Jan. 2018. doi: 10.3778/j.issn.1002-8331.1608-0081.

[124] M. Zhou, Q. Wang, J. Yang, Q. Li, F. Xiao, Z. Wang, and X. Chen, "PatternListener: Cracking Android pattern lock using acoustic signals," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Toronto, ON, Canada, 2018, pp. 1775–1787.

[125] Y. Zou, Q. Yang, Y. Han, D. Wang, J. Cao, and K. Wu, "AcouDigits: Enabling users to input digits in the air," in *Proc. IEEE PerCom*, Kyoto, Japan, Mar. 2019, pp. 1–9.

[126] Y. Zou, Q. Yang, R. Ruby, Y. Han, S. Wu, M. Li, and K. Wu, "EchoWrite: An acoustic-based finger input system without training," in *Proc. 39th IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Dallas, TX, USA, Jul. 2019, pp. 778–787.

[127] C. Yiallourides and P. P. Parada, "Low power ultrasonic gesture recognition for mobile handsets," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 2697–2701.

[128] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, New York, NY, USA, 2016, pp. 82–94.

[129] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, Niagara Falls, NY, USA, 2017, pp. 15–28.

[130] H. Chen, F. Li, and Y. Wang, "EchoTrack: Acoustic device-free hand tracking on smart phones," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun.*, Atlanta, GA, USA, May 2017, pp. 1–9.

[131] X. Xu, J. Yu, Y. Chen, Y. Zhu, and M. Li, "SteerTrack: Acoustic-based device-free steering tracking leveraging smartphones," in *Proc. 15th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Hong Kong, Jun. 2018, pp. 1–9.

[132] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "BatTracker: High precision infrastructure-free mobile device tracking in indoor environments," in *Proc. 15th ACM Conf. Embedded Netw. Sensor Syst.*, Delft, The Netherlands, 2017, Art. no. 13.

[133] D. Ensminger and L. J. Bond, *Ultrasonics: Fundamentals, Technologies, and Applications*, 3rd ed. New York, NY, USA: CRC Press, 2011.
[134] S. Lawoyin, X. Liu, D.-Y. Fei, and O. Bai, "Detection methods for a low-cost accelerometer-based approach for driver drowsiness detection," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, San Diego, CA, USA, Oct. 2014, pp. 1636–1641.
[135] C. Karatas, L. Liu, H. Li, J. Liu, Y. Wang, S. Tan, J. Yang, Y. Chen, M. Gruteser, and R. Martin, "Leveraging wearables for steering and driver tracking," in *Proc. IEEE INFOCOM-35th Annu. IEEE Int. Conf. Comput. Commun.*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.
[136] X. Xu, J. Yu, Y. Chen, Y. Zhu, and M. Li, "Leveraging acoustic signals for vehicle steering tracking with smartphones," *IEEE Trans. Mobile Comput.*, to be published. doi: 10.1109/TMC.2019.2900011.
[137] D. Tao, Y. Wen, and R. Hong, "Multicolumn bidirectional long short-term memory for mobile devices-based human activity recognition," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1124–1134, Dec. 2016.

ZHENGJIE WANG was born in Liaoyang, Liaoning, China, in 1972. He received the B.S. degree in power engineering from the North China University of Water Resources and Electric Power, Henan, China, in 1995, the M.S. degree in computer software and theory from Northeast University, Liaoning, China, in 2003, and the Ph.D. degree in computer application technology from the China University of Mining and Technology, Beijing, Beijing, China, in 2013.

Since 2003, he has been a Lecturer with the College of Electronic and Information Engineering, Shandong University of Science and Technology. He is the author of two books and more than ten articles. His research interests include human behavior recognition, people activity inference, person counting, identity authentication, and people tracking using WiFi devices and smartphones.

YUSHAN HOU was born in Jinan, Shandong, China, in 1995. She received the B.S. degree from the Shandong University of Science and Technology, in 2017, where she is currently pursuing the M.S. degree in communication and information system. Her research interests include deep learning, machine learning, and signal processing.

KANGKANG JIANG was born in Jining, Shandong, China, in 1993. She received the B.S. degree from Jining University, in 2017. She is currently pursuing the M.S. degree in communication and information system with the Shandong University of Science and Technology. Her research interests include image processing, machine learning, and deep learning.

WENWEN DOU was born in Taian, Shandong, China, in 1996. She received the B.S. degree from Weifang University, in 2018. She is currently pursuing the M.S. degree in communication and information system with the Shandong University of Science and Technology. Her research interests include deep learning, machine learning, and signal processing.

CHENGMING ZHANG was born in Zaozhuang, Shandong, China, in 1995. He received the B.S. degree from the Qilu University of Technology, in 2018. He is currently pursuing the M.S. degree in electronic and communication engineering with the Shandong University of Science and Technology. His research interests include machine learning, deep learning, and signal processing.

ZEHUA HUANG was born in Binzhou, Shandong, China, in 1996. He received the B.S. degree from the Shandong University of Science and Technology, in 2018, where he is currently pursuing the M.S. degree in communication and information system. His research interests include deep learning, machine learning, and signal processing.

YINJING GUO was born in Jining, Shandong, China, in 1966. He received the B.S. degree in radar engineering from the Ordnance Engineering College, in 1989, the M.S. degree in communication and electronic system, in 1992, and the Ph.D. degree in weapon system and application engineering from the Beijing Institute of Technology, in 2004.

From 1996 to 2019, he was a Professor with the College of Electronic and Information Engineering, Shandong University of Science and Technology. He has published more than 90 papers, among which more than 40 articles have been retrieved by the EI and SCI. His research interests include wireless communications, electromagnetic compatibility theory and applications, special radars, and unmanned aerial vehicle.

Dr. Guo has served as the Local Chair for the first, second, and third International Conference on Intelligent Information Technology Applications and the 3rd IEEE International Conference on Communication and Mobile Computing. He is a Reviewer of the National Natural Science Foundation of China, a Reviewer of the National Science and Technology Award, a member of the Qingdao Senior Experts Association, and a Reviewer for many international journals.