Amaging: Acoustic Hand Imaging for Self-adaptive Gesture Recognition

Penghao Wang[§], Ruobing Jiang[§] and Chao Liu[*]
Department of Computer Science and Technology, Ocean University of China, P.R.China
Email: wangpenghao@stu.ouc.edu.cn, jrb@ouc.edu.cn, liuchao@ouc.edu.cn

***Abstract*—A practical challenge common to state-of-the-art acoustic gesture recognition techniques is to adaptively respond to intended gestures rather than unintended motions during the real-time tracking on human motion flow. Besides, other disadvantages of under-expanded sensing space and vulnerability against mobile interference jointly impair the pervasiveness of acoustic sensing. Instead of struggling along the bottlenecked routine, we innovatively open up an independent sensing dimension of acoustic 2-D hand-shape imaging. We first deductively demonstrate the feasibility of acoustic imaging through multiple viewpoints dynamically generated by hand movement. *Amaging*, hand-shape imaging triggered gesture recognition, is then proposed to offer adaptive gesture responses. Digital Dechirp is novelly performed to largely reduce computational cost in demodulation and pulse compression. Mobile interference is filtered by Moving Target Indication. Multi-frame macro-scale imaging with Joint Time-Frequency Analysis is performed to eliminate image blur while maintaining adequate resolution. *Amaging* features revolutionary multiplicative expansion on sensing capability and dual dimensional parallelism for both hand-shape and gesture-trajectory recognition. Extensive experiments and simulations demonstrate *Amaging*'s distinguishing hand-shape imaging performance, independent from diverse hand movement and immune against mobile interference. 96% hand-shape recognition rate is achieved with ResNet18 and 60× augmentation rate.**

***Index Terms*—acoustic hand imaging; ubiquitous gesture recognition; mobile interference filtering; adaptive gesture response;**

I. Introduction

Motivation: Acoustic gesture recognition has received increasing attention due to the advantages offered by acoustic sensing, e.g., privacy-preserving, fast and low-cost deployment, higher accuracy and feasibility with low illumination or opaque obstacles. However, a significant issue in practice has been completely neglected by existing studies, i.e., self-adaptive segmentation to differentiate intended indicative motions from unintended transitional motions in natural human motion flow. As illustrated by Fig. 1, after the user performing a left swiping, a nature and necessary transition to withdraw the hand for the next left swiping will be rashly misidentified as a right swiping. A generally adopted but extremely awkward choice is to force users to perform unique marking gestures for manual segmentation. The failure to adaptively respond to meant gestures rather than unmeant transitions extremely reduces the ubiquitousness of acoustic gesture sensing.

Limitations of Prior work: Prior work on acoustic gesture recognition can be classified into two kinds, based on either trajectory tracking [1]–[9] or echo feature based pattern recognition [10]–[16]. The former kind tracks and recognizes target trajectory by roughly modeling the target hand/finger as a single reflecting particle. Such trajectory tracking approaches blindly keep tracking all through without adaptively differentiating meant gestures from unmeant transitions. Moreover, the simplified particle model is unable to filter out multi-path effect caused by complex hand shapes. To enhance the recognition resolution by taking the hand-shape factor into consideration, the later kind extracts integrated channel features from the reflected signals throughout each gesture. Due to the uninterpretable integrated features, where the 2-D hand-shape features cannot be independently extracted and recognized, the feature space and differentiating ability of such feature pattern based methods are limited while with high training cost.

Challenges: Based on a comprehensive insight into existing research advances, we have identified three advanced challenges to deploy practical ubiquitous gesture recognition.

1) **Adaptive response.** Ubiquitous gesture sensing requires smart identification of intended user gestures, while natural human movements will inevitably be interspersed with some unconscious transitional motions.
2) **Sensing capability expansion.** The feature space constructed by acoustic channel sensing has not been adequately developed, limiting the gesture differentiating capability.
3) **Mobile interference filtering.** The filtering of mobile

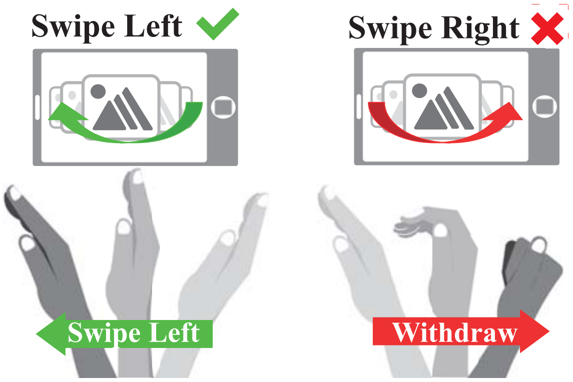


Fig. 1. Motivation of adaptive gesture recognition: an unintended but necessary withdraw preparing for the next gesture is unexpectedly responded.

[§]Co-first authors, [*]Corresponding author interference caused by close moving objects within the sensing range is still a significant open problem.

Our approach: Inspired by the recognition capability of human vision, we propose *Amaging* to enable acoustic signals 'see' 2-D hand shape while tracking the trajectory of a gesture. Only when an intended hand shape is seen/imaged, the segmentation will be adaptively triggered until the intended hand shape changes. To steadily image a moving hand, we transform the movement into an equivalent rotation and perform Rotating Objects Imaging. Multi-frame imaging is performed in macro time-scale by Joint Time-Frequency Analysis (JTFA) to eliminate target blur while maintaining adequate image resolution. *Amaging* filters out the interference from 'visible' mobile objects by Moving Target Indiction (MTI) filter. The significant advantage of *Amaging* is to multiplicatively expand the recognizable gesture space by opening up a new feature dimension. The multiplicative expansion is achieved to be the product of the scale of the two dimensions, hand-shape dimension and trajectory dimension.

Amaging can be implemented in off-the-shelf smartphones with speakers and microphones. The ratio of signal processing time to collection time is reduced lower than 1, enabling pipelined processing for further reduction on response delay. The major contributions are summarized as follows.

- We pioneer the new era of acoustic hand-shape imaging to establish a landmark in the area of acoustic gesture recognition. The proposed *Amaging* achieves a multiplicative expansion on acoustic gesture sensing capability by dimensional extension.
- *Amaging* steadily images the hand shape while moving, together with clarity improvement and $60\times$ data augmentation. The filtering of dynamic interference within the sensing range is also novelly addressed by MTI.
- The signal processing overhead has been greatly reduced by applying digital Dechirp, requiring single FFT for pulse compression. Extensive experiments demonstrate the distinguishing imaging performance irrelevant to movement, and 96% hand shape recognition rate.

II. The Big Picture

Acoustic motion sensing tracks time-varying 1-D radial distance of a moving target, e.g., hand and fingers. However, the routine to perform echo analysis has already become the bottleneck of current acoustic sensing techniques, severely limiting the sensing capability. As a result, we pioneer an independent sensing dimension of acoustic 2-D imaging in addition to existing available trajectory plotting [1]–[8].

In this section, we first explore the feasibility of applying acoustic signals for 2-D hand-shape imaging and briefly introduce the overview of the proposed *Amaging*. We then use the following three sections to describe three key components of *Amaging*. We evaluate the recognizing performance in Section VI, present the related work in Section VII and conclude in Section VIII.

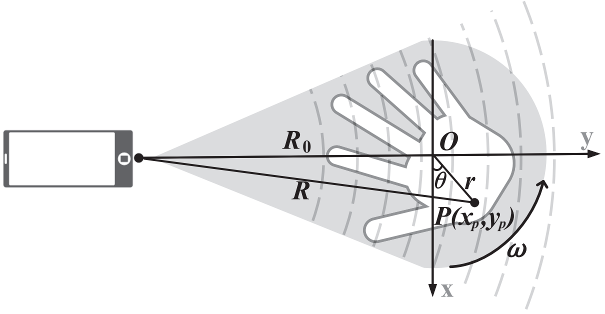


Fig. 2. Rotation based 2-D imaging model.

A. Feasibility Analysis

Whether the reflected echoes by a moving target contain enough information to portrait its 2-D image? Let's consider the visual experience that 3-D images can be visually captured through multiple viewpoints while single viewpoint only sees 2-D. As for acoustic sensing, a similar intuition arose: now that a single acoustic transceiver tracks the radial distance of all the reflectors on a moving target, multi-viewpoint transceivers will be able to image the 2-D shape of the target. Restricted by the fixed position (single viewpoint) of acoustic transceiver in practice, we turn to equivalently exploit target side rotation to dynamically generate multiple viewpoints. And fortunately, target rotation can be extracted from movement.

We now deductively demonstrate our intuition that target side rotation contributes to 2-D target-shape imaging. Fig. 2 illustrates an instantaneous projection of a moving target onto the plane of the instantaneous speed vector and the transceiver. Suppose the instantaneous movement of the target is decomposed into a rotation around O with an angular velocity ω. Any point P on the target projection with a distance r to O and an angle θ to x-axis has the coordinates

$$x_{P} = r\cos\theta, y_{P} = r\sin\theta. \tag{1}$$

The distance R between P and O at time t is then approximated as

$$\begin{aligned} R &= \sqrt{R_0^2 + r^2 + 2R_0 r\sin(\theta+\omega t)} \\ &\approx \sqrt{\left(R_0 + r\sin(\theta+\omega t)\right)^2} \\ &= R_0 + x_{P}\sin\omega t + y_{P}\cos\omega t \end{aligned} \tag{2}$$

where R_0 is the distance between the transceiver and O.

The Doppler Frequency Shift (DFS) f_d of the echo reflected by P is

$$f_d = \frac{2}{\lambda}\frac{dR}{dt} = \omega\left(\frac{2x_{P}}{\lambda}\cos\omega t + \frac{2y_{P}}{\lambda}\sin\omega t\right), \tag{3}$$

where λ is the acoustic wavelength. When $t \to 0$, the doppler shift f_d and round-trip propagation delay τ, which can be

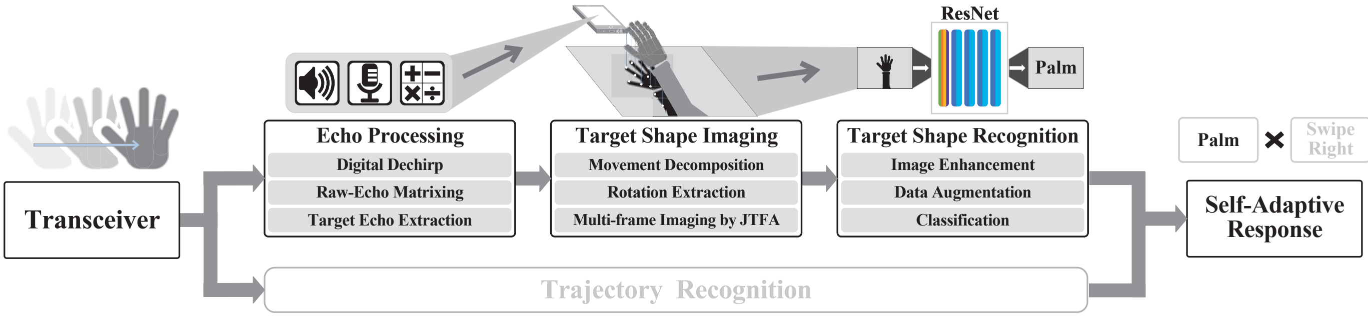


Fig. 3. System overview of the proposed *A*coustic target-shape i*maging* based gesture recognition, named *Amaging*.

detected by the transceiver through echo analysis, can be expressed by

$$\begin{aligned} f_d &= \frac{2\omega}{\lambda} x_P, \\ \tau = \frac{2R}{c} &= \frac{2}{c} y_P + \frac{2R}{c}. \end{aligned} \tag{4}$$

Thus the 2-D coordinates (x_P, y_P) of P can be achieved.

As a result, it is feasible to acoustically image real-time 2-D projections of a moving target by extracting real-time rotations from the movement. Note that the real-time projection plane is determined by the instantaneous line-of-sight vector and the instantaneous moving vector.

B. System Overview

An orthogonal sensing dimension of target shape imaging has been built up by the proposed *Amaging* in addition to existing target trajectory recognition. *Amaging* expands the gesture recognition space through multiplying the scale of target-shape dimension by the scale of target-trajectory dimension.

The outline of *Amaging*, illustrated in Fig. 3, is to process the raw echoes, image and recognize the target shape, simultaneously recognize target trajectory, and adaptively respond to intended gestures.

Echo Processing (Section III): Digital Dechirp performs demodulation and primary pulse compression on the received raw echo. Then the raw echoes of the sequentially sent symbols are aligned into matrices to present the distance of all reflectors in a macro time-scale. Thus the movement evolution can be tracked within a second-level time window, instead of symbol-by-symbol. The sensing area will be dynamically focused around the target, e.g., the hand, indicated by the reflectors with relatively greater Doppler Frequency Shift (DFS). The slower and static reflectors, e.g., the arm, can finally be filtered out with MTI filter.

Target Shape Imaging (Section IV): As analyzed in Section II-A, it is feasible to keep imaging a moving target in real time by extracting the rotation component of each instantaneous movement. We first model the measurable responses of echo parameters to movement parameters. The formulated echo responses is further decomposed into rotational and translational component. The later part is then compensated by using Image Contrast Based Autofocus (ICBA) to estimate translation parameters, e.g., radial velocity. Finally, each symbol matrix is multi-frame imaged by JTFA to generate an image matrix, instead of a single image, to avoid target blur and increase original samples.

Target Shape Recognition (Section V-A): Images from each image matrix are enhanced in batches by centering target and jointly denoising. Data augmentation is performed based on originally sampled frames to improve training efficiency. ResNet18 is used for hand shape classification.

Self-adaptive Response (Section V-B): *Amaging* is enabled to respond as expected to exactly those enrolled gestures scattered in a coherent motion flow. The response mechanism is based on parallel recognition along two orthogonal gesture dimensions. The target-shape dimension triggers the target-trajectory dimension and indicates the corresponding segment of an intended gesture. The segment range determination based on image-sequence joint analysis is fault tolerant and robust against burst interference.

III. Echo Processing

A. Digital Dechirp

Digital Dechirp is innovatively proposed to migrate the advantages of analog Dechirp to portable-device scenarios. The advantage of Dechirp to perform pulse compression over time/frequency-domain matched filtering is the significant reduction in computational cost and intermediate storage. Dechirp applies a signal mixer for both demodulation and pulse compression, only requiring signal multiplications and once FFT. Whereas existing time-domain matched filtering [17]–[20] requires the compute-intensive convolution on signals, and frequency-domain matched filtering [21], [22] requires multiple FFT/IFFT operations and considerable storage space.

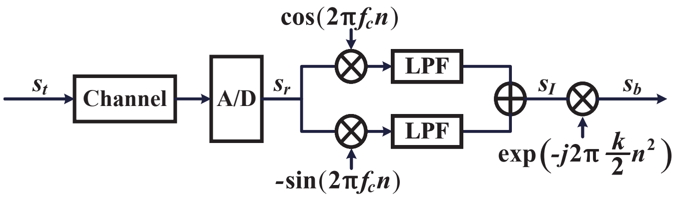


Fig. 4. Flow chart of digital Dechirp

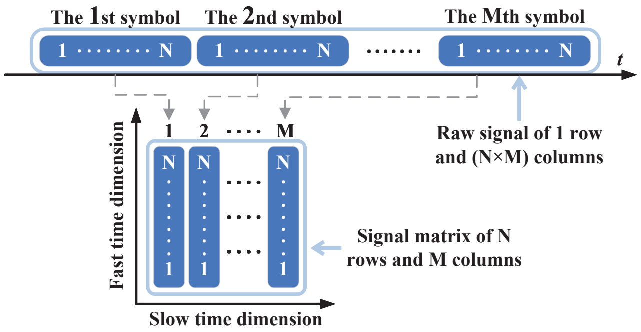


Fig. 5. Echo matrixing

To address the issue that the signal mixer used in analog Dechirp is not available in existing portable devices, we novelly perform digital-manner Dechirp, integrating IQ demodulation and pulse compression after A/D conversion, illustrated by Fig. 4. For the transmitted signal

$$
\begin{aligned}
s_t\left(t\right) &= \mathrm{rect}\left(\frac{t}{T_t}\right)\cos\left(j2\pi\left(f_c+\frac{k}{2}t\right)t\right),\\
\mathrm{rect}(\cdot) &= \begin{cases}1, & -\frac{1}{2}\leq\cdot\leq\frac{1}{2}\\ 0, & \text{rest}\end{cases},
\end{aligned}
\tag{5}
$$

where f_c is the carrier frequency, k is the chirp rate and T_t is the transmitting duration within a whole symbol duration T, the received signal after A/D conversion is

$$
s_r(n) = \mathrm{rect}\left(\frac{n-\tau}{T_t}\right)\cos\left(2\pi\left(f_c+\frac{k}{2}(n-\tau)\right)(n-\tau)\right), \tag{6}
$$

where n is often short for nT_s and T_s is the sampling period. The Intermediate Frequency(IF) digital signal, $s_I(n)$, is achieved by two-way multiplication, low-pass filtering, and combining,

$$
s_I\left(n\right) = \mathrm{rect}\left(\frac{n-\tau}{T_t}\right)\exp\left(-j2\pi f_c\tau+j\pi k\left(n-\tau\right)^2\right). \tag{7}
$$

The digital mixing is thus completed to achieve the base signal $s_b(n)$ by multiplying $s_I(n)$ with the conjugate of the reference signal $s_{\mathrm{ref}}(n)=\exp\left(j\pi kn^2\right)$,

$$
s_b\left(n\right) = \mathrm{rect}\left(\frac{n-\tau}{T_t}\right)\exp\left(-j2\pi f_c\tau-j2\pi k\tau n+j\pi k\tau^2\right). \tag{8}
$$

The pulse compression will be efficiently executed with single FFT on s_b in later matrixing (Section III-B) and imaging (Section IV-C) for direct signal localization and range profile generation, respectively, i.e.,

$$
\begin{aligned}
S_b\left(f\right) = &T_t\mathrm{sinc}(\pi T_t\left(f+k\tau\right))\exp\left(-j2\pi f\tau\right)\\
&\exp\left(-j2\pi f_c\tau\right)\exp\left(j\pi k\tau^2\right).
\end{aligned}
\tag{9}
$$

B. Echo Matrixing

The received signals are aligned to the direct arrivals of sent symbols since the direct path brings obvious largest signal amplitude. As shown in Fig. 5, starting with the direct arrived signal, each symbol has N sampling periods T_s and every M consecutive symbols form a symbol matrix of $(N\times M)T_s$.

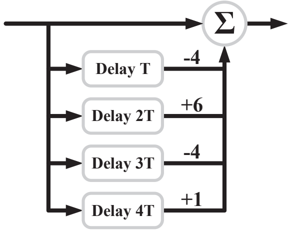


Fig. 6. MTI structure.

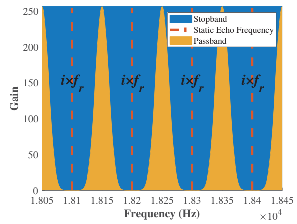


Fig. 7. Power gain of MTI.

Identification of direct signal. By performing pulse compression on a segment of received signals, signal delay and the corresponding amplitude can be achieved. The delay with largest amplitude will be considered as direct arrivals. Note that the pulse compression can be efficiently performed with our digital Dechirp and only once FFT on the base signal s_b.

Basics of echo Matrix. According to the different speeds of time change along the two directions, we name the horizontal direction as slow-time dimension and the vertical direction as fast-time dimension. The corresponding 2-D base signal to the 2-D signal matrix, denoted by $s_b(n,m)$, can be obtained by performing parallel digital Dechirp on M aligned symbols along the fast time dimension. The 2-D range profile $\mathbf{R}(n,m)=\mathrm{FFT}(s_b(n,m),1)$ corresponding to the 2-D signal matrix can also be efficiently obtained for further 2-D target shape imaging in Section IV-C.

C. Target Echo Extraction

We extract target echoes by filtering out not only static echoes but also slower moving object echoes in presence of random noises. Existing sliding-window subtraction method [23] fails for stable elimination of static echoes with spatio-temporal random noise, let alone dynamic echoes. Instead of struggling in time domain, we novelly shift to frequency-domain distinguishing on diverse DFS.

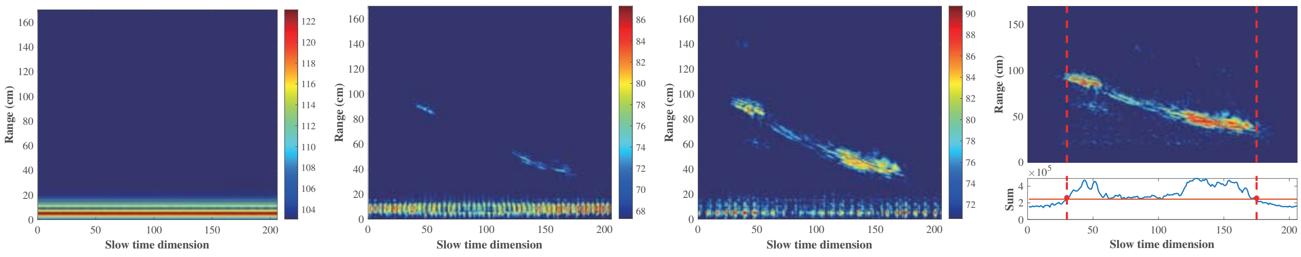
We apply Moving Target Indiction (MTI) filter [24], a frequency-domain combo band-pass filter, to automatically filter out low-DFS static and dynamic echoes. For static echoes and direct signals without DFS, such periodic signals with the pulse repetition frequency (PRF) $f_r = 1/T$ can be broken down into harmonics with integer multiples of f_r, i.e., $if_r, i=1,2,\ldots$. While target echoes with greatest DFS have deviated frequency if_r+f_d. The structure of the applied MTI, presented in Fig. 6, is thus designed to stop signals with $f=if_r$ and pass signals with $f=if_r+f_d$. The impulse response can be expressed as,

$$
\begin{aligned}
h\left(t\right) = &\delta\left(t\right)-4\delta\left(t-T\right)+6\delta\left(t-2T\right)-4\delta\left(t-3T\right)\\
&+\delta\left(t-4T\right).
\end{aligned}
\tag{10}
$$

The corresponding power gain illustrated in Fig. 7 is

$$
\left|H\left(f\right)\right|^2 = 256\left(\sin\left(2\pi fT/2\right)\right)^8. \tag{11}
$$

As to slower moving objects with small DFS, the smaller their DFS, the closer their frequency with if_r, which will also be filtered. According to Equ. (11), signals with $f\in$



(a)

(b)

(c)

(d)

Fig. 8. Performance comparison of target echo extraction by sliding-window subtraction and MTI filter. (a) $\mathbf{R}(n,m)$ of raw symbol matrix. (b) $\mathbf{R}(n,m)$ with sliding-window subtraction. (c) $\mathbf{R}(n,m)$ with MTI filter. (d) Dynamically target range focusing.

$((i+0.167)f_r, (i+0.833)f_r)$ are largely retained. Suppose $f_r = 100\,\mathrm{Hz}$, movements within the speed range of 16.7 cm/s - 83 cm/s will pass.

The combo-shaped MTI filter can be elegantly implemented by performing weighted sum on corresponding columns of our echo matrices. Fig. 8 shows the target extraction performance of MTI filter in the form of 2-D range profile over existing time-domain sliding-window subtraction. In the experiment, the target hand pushes towards the smartphone transceiver. Fig. 8(a) shows the range profile $\mathbf{R}(n,m) = \mathrm{FFT}(s_b(n,m),1)$ of a raw symbol matrix. Only static echoes and direct signals with much larger amplitude over the target echoes can be obviously seen. Fig. 8(b) presents the resulted $\mathbf{R}(n,m)$ with time-domain subtraction, where the target is not obvious due to vast residual static echoes and direct signals with random noises. Fig.8(c) illustrates the distinguishing performance of MTI filter. The pushing from around 95 cm to around 35 cm during about 150 units along the slow-time dimension (1.5 s) can be obviously observed. The target extraction is further enhanced by dynamic target range focusing. As shown in Fig. 8(d), the residual clutters much closer than the target range is first removed. Then the target range along the other slow-time dimension is also automatically determined by setting a dynamic signal amplitude threshold,

$$\mathrm{Th} = \mathrm{Avg} + k \cdot \mathrm{Std}, \tag{12}$$

where Avg and Std are respectively the mean and standard deviation of the vertical accumulations on the range profile, i.e., $\mathrm{Sum}(\mathbf{R}(n,m),1)$, and k is a customized parameter.

IV. Target Shape Imaging

A. Movement Decomposition

The 2-D base signal $s_b(n,m)$ is modeled as separate impacts of decomposed movement components on signal parameters. For a matriculated 2-D sent signal,

$$s_t(n,m) = \exp\left(j2\pi\left(f_0 + \frac{k}{2}n\right)n\right)\mathrm{rect}\left(\frac{n-mT}{T_t}\right), \tag{13}$$

the reflected signal s_r from a reflector $\boldsymbol{X} = (x,y,z)$ can be received as

$$\begin{aligned} &s_r(\boldsymbol{X},n,m) = \rho(\boldsymbol{X})s_t(n-\tau(\boldsymbol{X},n,m),m), \text{ where} \\ &\tau(\boldsymbol{X},n,m) = 2R(\boldsymbol{X},n,m)/c \end{aligned} \tag{14}$$

denotes the echo delay, $\rho(\boldsymbol{X})$ denotes the signal reflection intensity of $\boldsymbol{X}$, and $R(\boldsymbol{X},n,m)$ is the absolute distance of $\boldsymbol{X}$ with respect to the transceiver at time $[(m-1)N+n]T_s$.

The received signal integrating all the echoes from all reflectors over the whole target can be achieved as,

$$s_r(n,m) = \int \rho(\boldsymbol{X})\, s_t(n-\tau(\boldsymbol{X},n,m),m)\,\mathrm{d}\boldsymbol{X}. \tag{15}$$

As analyzed in Section II-A, the distance R of a reflector $\boldsymbol{X}$ can be transformed into the absolute distance of a reference point O combining the relative distance of $\boldsymbol{X}$ towards O, i.e.,

$$R(\boldsymbol{X},n,m) \approx R_0(n,m) + \boldsymbol{x}\cdot\boldsymbol{i}_{\mathrm{LoS}}(n,m), \tag{16}$$

where $R_0(n,m)$ is the reference distance, $\boldsymbol{x} = (x',y',z')$ is the relative coordinate of $\boldsymbol{X}$ to the reference point O in the reference coordinate system, and $\boldsymbol{i}_{\mathrm{LoS}}(n,m)$ denotes the unit vector along the line of sight direction. Thus the delay in Equ. (15) is converted to

$$\tau(\boldsymbol{X},n,m) = \tau_0(n,m) + \tau'(\boldsymbol{x},n,m), \tag{17}$$

where $\tau_0(n,m) = 2R_0(n,m)/c$ denotes the delay change due to target translation, i.e., the radial distance change of O, and $\tau'(\boldsymbol{x},n,m) = 2\boldsymbol{x}\cdot\boldsymbol{i}_{\mathrm{LoS}}(n,m)/c$ denotes the delay change due to the relative rotation of $\boldsymbol{x}$ towards O. The received signal can be finally achieved as,

$$s_r(n,m) = \int \rho(\boldsymbol{x})s_t(n-\tau_0(n,m)-\tau'(\boldsymbol{x},n,m),m)\,\mathrm{d}\boldsymbol{x}. \tag{18}$$

The corresponding 2-D base signal is also obtained,

$$\begin{aligned} s_b(n,m) &= \exp(j\varphi_0(n,m)) \\ &\int \rho(\boldsymbol{x})\exp(j(\varphi_1(\boldsymbol{x},n,m)+\varphi_2(\boldsymbol{x},n,m)))\,\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{19}$$

where

$$\begin{aligned} \varphi_0(n,m) &= -2\pi(f_c+kn)\cdot\tau_0(n,m) + \pi k\tau_0^2(n,m), \\ \varphi_1(\boldsymbol{x},n,m) &= -2\pi(f_c+kn)\cdot\tau'(\boldsymbol{x},n,m), \\ \varphi_2(\boldsymbol{x},n,m) &= -2\pi k\tau_0(n,m)\cdot\tau'(\boldsymbol{x},n,m) \\ &\quad + \pi k(\tau'(\boldsymbol{x},n,m))^2, \end{aligned} \tag{20}$$

φ_0 is the phase change caused by target translation and φ_1 and φ_2 are the phase changes caused by relative rotation.

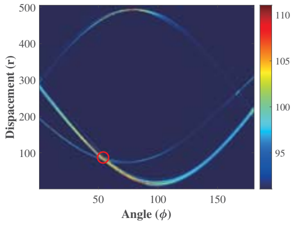


Fig. 9. $\mathbf{RT}(\phi, \mathbf{r})$.

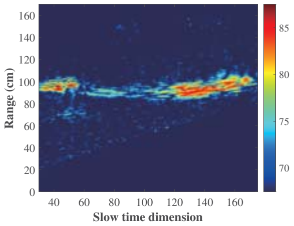


Fig. 10. Compensate Fig. 8(d).

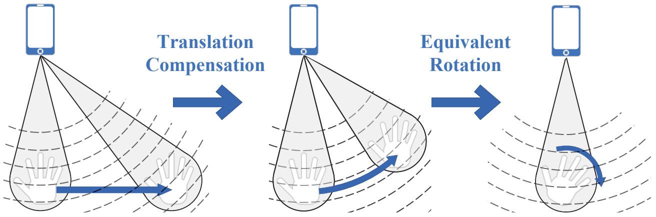


Fig. 11. Illustration of translation compensation and rotation extraction.

B. Rotation Extraction

The rotational effect on the base signal s_b is extracted by compensating s_b with a compensation signal to be the conjugate of the translational effect, i.e.,

$$s_b\left(n, m\right) \cdot \exp(-j\varphi_0(n, m)). \tag{21}$$

We now generate the compensation signal by estimating φ_0 in following two steps.

1) Modeling the translational effects: The absolute distance of the reference point during a short time can be expressed as

$$R_0\left(t\right) = R_0 + v_{_R} t + \frac{1}{2} a_{_R} t^2, \tag{22}$$

where R_0 is the initial distance, $v_{_R}$ and $a_{_R}$ are the radial velocity and acceleration of the reference point, respectively. Since the acoustic signals are transmitted by symbols with the period $T = 10\,\text{ms}$, the initial distance of the m th symbol can be expressed as

$$R_0\left(m\right) = R_0 + v_{_R} mT + \frac{1}{2} a_{_R}\left(mT\right)^2. \tag{23}$$

During the extremely short symbol period, target velocity within each symbol can be approximated to be stable, i.e.,

$$v_{_R}(m) = v_{_R} + a_{_R} mT. \tag{24}$$

Thus the target distance during m th symbol is obtained,

$$R_0\left(n, m\right) = R_0\left(m\right) + v_{_R}\left(m\right) n. \tag{25}$$

Finally the translation effect on echo phase, $\varphi_0\left(n, m\right)$, can be formulated as

$$\begin{aligned}\varphi_0\left(n, m\right) = & -2\pi\left(f_0 + kn\right)\left(\tau_0\left(m\right) + \frac{2}{c} v_{_R}\left(m\right) n\right) \\ & + \pi k\left(\tau_0\left(m\right) + \frac{2}{c} v_{_R}\left(m\right) n\right)^2,\end{aligned} \tag{26}$$

where $\tau_0(m) = 2R_0(m)/c$. In summary, φ_0 is actually a function of $v_{_R}, a_{_R}$. Therefore, it is feasible to compensate φ_0 with the radial movement parameters $v_{_R}, a_{_R}$.

2) Estimating Translation Parameters by ICBA: The radial velocity $v_{_R}$ is first estimated with Radon Transform [25] on target range profile.

The radial velocity $v_{_R}$ is actually the slope of the target distance trajectory, which can be observed from the range profile shown in Fig. 8(d), i.e., $v_{_R} = \tan\left(\phi\right)$. ϕ is the angle between the distance trajectory and the time dimension, which can be determined by Radon Transform on $\mathbf{R}(n, m)$,

$$\hat{\phi} = \arg\left(\max_\phi(\mathbf{RT}\left(\phi, r\right))\right) - \frac{\pi}{2}, \tag{27}$$

where $\mathbf{RT}\left(\phi, r\right)$ in Fig.9 is the results of Radon Transform on Fig.8(d). The estimated angle $\hat{\phi}$ is the x-coordinate of the circled point and the velocity is estimated as $\widehat{v}_{_R} = \tan\left(\hat{\phi}\right)$.

The radial acceleration $a_{_R}$ is then estimated by linear search and Image Contrast Maximization. Potential accelerations $\widetilde{a}_{_R} \in \left[\min\left(a_{_R}\right), \max\left(a_{_R}\right)\right]$ are sequentially selected to compensate the base signal s_b along with $\hat{v}_{_R}$. Then imaging trials with FFT2 are performed for test,

$$\mathbf{I}\left(\hat{v}_{_R}, \widetilde{a}_{_R}\right) = \text{FFT2}\left(s_b\left(n, m\right) \cdot \exp\left(-j\varphi_0(n, m, \hat{v}_{_R}, \widetilde{a}_{_R})\right)\right). \tag{28}$$

The following metric of Image Contrast(IC) is maximized over all the potential accelerations,

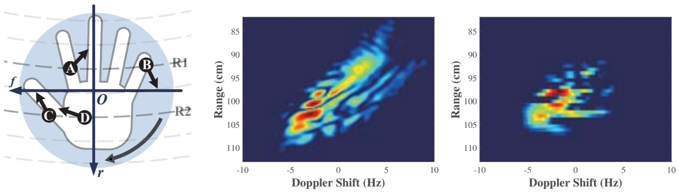
$$\mathbf{IC}\left(\hat{v}_{_R}, \widetilde{a}_{_R}\right) = \frac{\sqrt{\text{Mean}\left(\mathbf{I}\left(\hat{v}_{_R}, \widetilde{a}_{_R}\right) - \text{Mean}\left(\mathbf{I}\left(\hat{v}_{_R}, \widetilde{a}_{_R}\right)\right)\right)}}{\text{Mean}\left(\mathbf{I}\left(\hat{v}_{_R}, \widetilde{a}_{_R}\right)\right)}. \tag{29}$$

The metric IC indicates the focus degree of imaging. Since the focus degree essentially depends on the compensation performance of $\widetilde{a}_{_R}$, the best acceleration can be determined as the one maximizing $\mathbf{IC}\left(\hat{v}_{_R}, \widetilde{a}_{_R}\right)$, i.e.,

$$\hat{a}_{_R} = \arg\left(\max_{\widetilde{a}_{_R}}\left(\mathbf{IC}\left(\hat{v}_{_R}, \widetilde{a}_{_R}\right)\right)\right). \tag{30}$$

Now we can use the translation parameters $\hat{v}_{_R}$ and $\hat{a}_{_R}$ to compensate target radial movement for rotation extraction. Fig. 10 exhibits the compensation performance on the range profile in Fig. 8(d). It is obvious in the figure that the target radial distance along the time dimension has been flattened at a stable distance around 95 cm.

As illustrated in Fig. 11, the translation compensation to eliminate radial movement component is equivalent to pull the target moving at a fixed distance. Thus multiple viewpoints can be dynamically generated for 2-D shape imaging with the equivalent rotation model demonstrated in Section II-A.



(a) Rotating palm (b) Basic imaging (c) Advanced imaging

Fig. 12. Distinguishing imaging performance of the advanced imaging over the basic imaging.

C. Multi-farme Imaging by JTFA

Basic Imaging. A basic rotation imaging model can be constructed based on the compensated base signal,

$$\begin{aligned} s_b'(n,m) &= s_b(n,m)\cdot\exp\left(-j\hat{\varphi}_0(n,m)\right) \\ &= \int \rho'(x,y)\exp\left(j\left(x\cdot Q_1 + y\cdot Q_2\right)\right)\mathrm{d}x\,\mathrm{d}y, \\ Q_1 =& \frac{2}{c}\left(f_c - \tau_0(m) - \frac{1}{2}\tau'(x,y,n,m)\right)\cos\left(\theta(n,m)\right), \\ Q_2 =& \frac{2}{c}\left(f_c - \tau_0(m) - \frac{1}{2}\tau'(x,y,n,m)\right)\sin\left(\theta(n,m)\right), \end{aligned} \tag{31}$$

where $\rho'(x,y)$ is the reflection intensity by the projection of $\boldsymbol{x}$ onto the plane containing target moving vector and orthogonal to the rotation axis. A single-frame 2-D target shape image is achieved by performing dual-dimension FFTs on $s_b'(n,m)$,

$$\mathbf{I}(n_r,m_f) = \mathrm{FFT2}\left(s_b'(n,m)\right) = \mathrm{FFT}\left(\mathbf{R}(n,m),2\right), \tag{32}$$

where $\mathbf{I}(n_r,m_f)$ represents the reflectivity of the target projection. $\mathrm{FFT2}(\cdot)$ denotes dual-dimension FFTs, the first along the fast time dimension and the second along the slow time dimension. The first dimension FFTs produce range profile $\mathbf{R}(n,m)$, representing the distance image of all the reflectors on the target projection. The second dimension FFTs, $\mathrm{FFT}\left(\mathbf{R}(n,m),2\right)$, produce distance-dimension spectrum of same-distance reflectors over different distances. Different reflectors at same distance can be distinguished in the reflectivity image because they have different Doppler shifts in the spectrum during rotation. As in Fig. 12(a), A, B at distance R_1 and C, D at distance R_2 will be separately ranged by first dimensional FFTs. A with positive DFS and B with negative DFS will further be separated by zero-f axis. C and D with same range and both positive DFS but different DFS magnitude will accordingly be distinguished.

Advanced Imaging. The basic imaging, producing single image for an echo matrix, has the significant disadvantage of image blur. In the rotation imaging model, the symbols spreading a large time scale in a matrix record non-negligible target rotation angle. Thus the overlapping of rotating images results in the imaging blur. Unfortunately, simply shortening the echo matrix will reduce the image resolution. We address the image blur issue by applying Joint Time-Frequency Analysis (JTFA) [26], [27] to replace $\mathrm{FFT}\left(\mathbf{R}(n,m),2\right)$. Overlapped multiple frames are separately imaged from single matrix to improve clarity. Moreover, the original image sample set is largely increased to reduce the cost of training data collection. Fig. 12 demonstrates the obvious higher clarity of JTFA-based imaging over FFT-based option.

V. Imaging based Adaptive Response

A. Image Enhancement and Target Shape Recognition

The imaged multiple frames of spectrum-range reflectivity $\mathbf{I}(n_r,m_f)$ are further enhanced by image processing to facilitate target shape recognition.

Target centering and noise eliminating. *Amaging* supports dynamic gesture sensing range by automatic target range focusing. Thus targets may locate at diverse ranges in the reflectivity image. We center targets by uniformly shifting the max-reflectivity point to the zero-range axis. In addition, random noises are eliminated by setting a dynamic reflectivity threshold as Equ. (12).

Data augmentation. In addition to the multi-frame increase of original image samples (5× augmentation rate) in Section IV-C, a series of data augmentation techniques are applied to reduce data collection cost.

- Horizontal scaling (3× augmentation rate): Original images are horizontally scaled to emulate non-uniform hand moving velocity.
- Mirroring (2× augmentation rate): By horizontally flipping 2-D hand-shape images, unilateral gestures can be doubled to bilateral gestures. Bilateral gesture recognition thus is elegantly available.
- Salt-Pepper Noise (2× augmentation rate): To prevent overfitting and increase recognition robustness, salt-pepper noises are added.

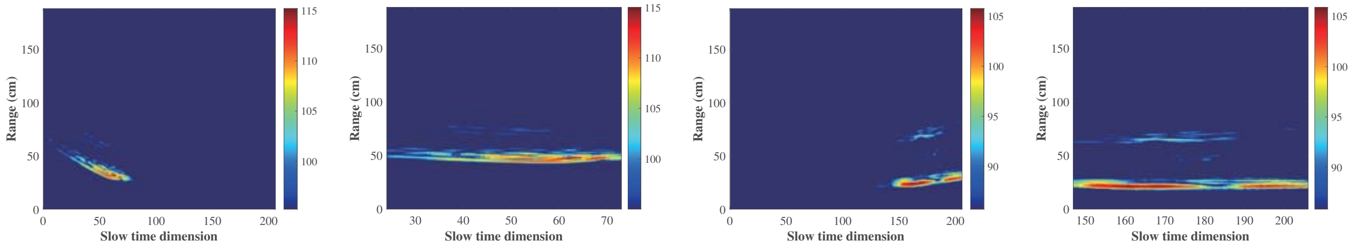
Classification. The desired collection of enrolled hand shapes can be customized. The training network optimization is out of the research scope and thus a general ResNet [28] is used as our classifier. Note that our gesture response mechanism features error-correction capability against possible classification faults of general classifiers.

B. Adaptive Response Mechanism

Amaging responses to enrolled gestures with shape & trajectory double dimensions among the flow of natural human motions. The hand-shape image sequence is jointly analyzed and adaptively segmented, which triggers the trajectory-dimension recognition on the corresponding section of the symbol stream. Classification faults or burst interference, can be corrected by a sliding window on the image sequence to jointly vote the local dominant shape. Then the frame range of a gesture can be immediately determined once the dominant shape changes. The continuous hand shape imaging and trajectory recognition on selective sections can be highly paralleled to reduce response delay.

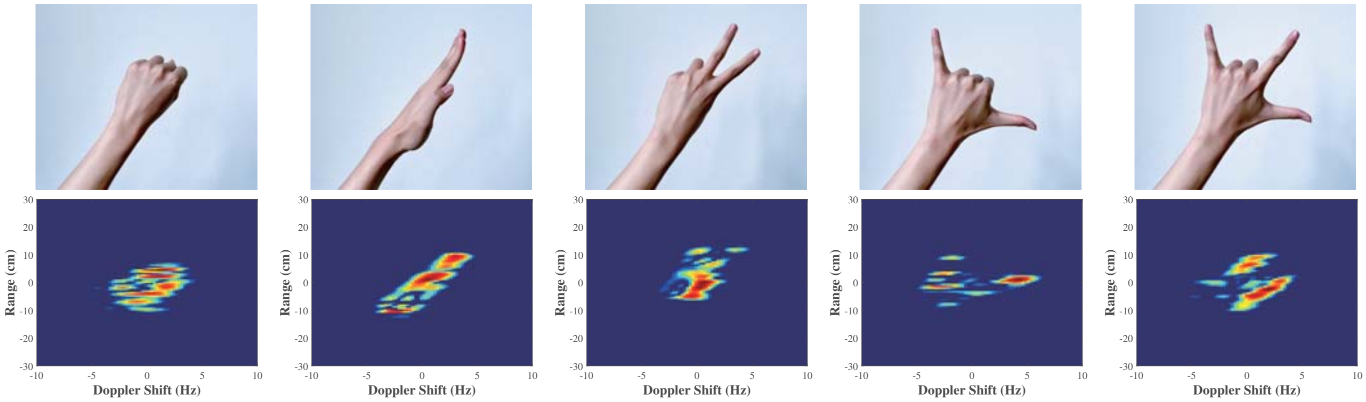
VI. Performance Evaluation

In this section, we first demonstrate the orthogonality of the novel target shape dimension against the trajectory dimension commonly used by state-of-the-art studies. Since the integrated gesture recognition rate, affected by both dimensions, is incapable to reflect the shape-dimension recognition performance,



(a) Fast moving palm (b) Rotation extraction on (a) (c) Slowly moving palm (d) Rotation extraction on (c)

Fig. 13. Independent rotation extraction performance to distinct movements.



(a) Rock (b) Paper (c) Scissors (d) Six (e) Love

Fig. 14. Imaging performance of common gestures.

we individually exhibit the shape imaging performance and evaluate the joint shape recognition accuracy.

A. Setup

Experimental parameters. Off-the-shelf smartphone Samsung Note10 Plus is used to transmit and receive linear frequency modulated (LFM) signals from 10 kHz to 24 kHz [29]. Symbol period $T = 10\,\text{ms}$. Transmission duration $T_t = 5\,\text{ms}$. Sampling rate $f_s = 48\,\text{Hz}$.

Data collection and training. We invited 10 volunteers (5 men and 5 women), two of whom are left-handers, to perform single-hand gestures of the 6 hand shapes with different speed and trajectories. Training samples are collected from 8 volunteers, each conducting 10 samples per hand shape. The final number of samples is 28,800 with 60× data augmentation rate. Testing samples are collected from the other 2 volunteers. Each sample is conducted within 3 s. 4 experimental scenarios are varied in terms of different background objects and noise intensities. ResNet18 is selected to be the classifier, trained by PyTorch on a server with Intel Core i9-11900K CPU, 32 GB RAM, and a GeForce RTX 3090 Graphics Card.

B. Dimensional Independence Demonstration

We first exhibit the independent shape imaging performance exhausting the movement dimension. The imaging performance is represented by the dominant factor of rotation extraction performance.

Each hand shape is repeated with extremely different speeds, start and end positions, and start time. The rotation extraction performance for the hand shape with greatest moving difference are exhibited by Fig. 13. Fig. 13(a) and Fig. 13(c) respectively show a palm moving from far right to near left early and rapidly, and from near left to far right late and slowly. The respective rotation extraction results are given by Fig. 13(b) and Fig. 13(d). It is obvious that movement does not impact the rotation model of a hand shape. In other words, the imaging performance is independent of hand movements.

In addition, the experiments are conducted with a moving interference at a comparable speed with the slower moving in Fig. 13(c) at the range of 70 cm. The interference is more obvious in Fig. 13(c) than in Fig. 13(a) because of the different speed gaps. The robustness of *Amaging* against such mobile interference will next be shown by the imaging performance.

C. Imaging Performance and Joint Recognition Accuracy

The imaging performance on the other 5 hand shapes are shown in Fig. 14 except the palm in Fig. 12. The shapes can be easily distinguished. The fragmentary image is caused by uneven reflectivity over the hand surface.

Separate Recognition Rate. Fig. 15 presents the confusion matrix of average recognition rate of six hand shapes with 100 test frames (20 test conductions) for each shape. From the figure, Rock and Paper have 100% recognition rate, while other hand shapes have a fault rate lower than 6% because of

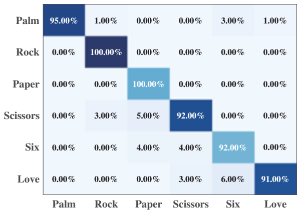


Fig. 15. Confusion matrix of average recognition rate over six kinds of hand shapes.

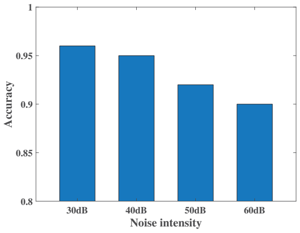


Fig. 16. Shape recognition rate with joint analysis vs. noise intensity

the shape similarity. For example, Love may be recognized as Six with 6% probability.

Joint Recognition Rate. In this experiment, fault correction is applied on multiple images of each test trial by jointly voting the dominant shape. In noisy scenarios, the average recognition rates over all the hand shapes with 100 tests for each are shown in Fig. 16. *Amaging* achieves distinguishing recognition rate, 96% with 30 db audible sound.

Response Delay. The simulated running time to process a matrix containing 100 symbols (lasting 1 s) is presented in Tab. I. The pre-processing operations can be approximately finished with the current symbol collection and the processing of current matrix (0.917 s) can be finished during the next round (1 s). Thus pipe-lined processing can be enabled to further reduce response delay.

VII. Related Work

State-of-the-art acoustic gesture recognition techniques can be classified into trajectory-based ones [1]–[8] and echo feature pattern based ones [10]–[16]. In addition, the only acoustic imaging attempt with lightweight devices [29] is finally be introduced.

A. Trajectory based Recognition

This kind of recognition methods exploit single dimension of gesture trajectory, neglecting the parallel hand-shape dimension with simplified particle model. The major challenge for the particle model is the filtering of multi-path effect [30] caused by actually complex hand shapes. This category further develops along two different directions.

Relative displacement accumulation. Target initial positions are needed to be determined, uninterruptible tracking is required and accumulative errors may occur. Representative solutions are LLAP [1] and Strata [2]. LLAP detects echo phase changes reflecting target displacement. Strata performs time-domain channel estimation to improve tracking accuracy.

Absolute distance measurement. Round-trip Time of Flight (ToF) is commonly used to range target absolute distance to avoid accumulative error. FingerIO [3] exploits phase encoded OFDM for ToF measurement. AMT [4] tracks multiple targets by constructing a MIMO system with speaker-mic pairs.

B. Echo Feature Pattern based Recognition

This kind of approaches learn echo features reflecting the integrated impacts of both hand shape and movement, which is a good attempt to extend gesture features. However, the two feature dimensions are not independently considered and the uninterpretable features severely limit the differentiation space. In addition, considerable training cost based on large scale training data collection is introduced.

TABLE I
Amaging's processing delay on 100 symbols

Pre-processing			**Processing**		
Matrixing	MTI	DDechirp	ICBA	Imaging	Classification
0.004s	0.003s	0.153s	0.683s	0.152s	0.082s

AudioGest [10] only learns the Doppler frequency shift reflecting radial velocity of moving targets to recognize six gestures. RobuCIR [12] exploits Channel Impulse Response (CIR) to track real-time 1-D radial distance of a moving target. Combining CNN and LSTM for image feature extraction and gesture recognition, 15 gestures can be differentiated.

C. Acoustic Imaging by Lightweight devices

Acoustic target-shape imaging opens up a new sensing dimension to develop acoustic sensing capability. Combining both radial distance tracking and Doppler frequency shift differentiation at same range, all the reflectors on the target can be dual-dimensional imaged. Image recognition algorithms are also required for target shape classification.

AIM [29] is the only work on target-shape imaging with lightweight devices such as smartphones in recent years. However, AIM cannot be used for gesture recognition. Firstly, the imaged target is required to keep static during the imaging. Secondly, the imaging device is required to dynamically scan the whole fixed target. Finally, the relative movement between the device and the target should be known, including scan distance and moving trajectory.

VIII. Conclusion

The novel dimension of 2-D target-shape imaging has been opened up for acoustic sensing by this work. Targeting at the identified three key challenges on ubiquitous acoustic gesture recognition, namely adaptive response, sensing capability expansion, and mobile interference filtering, a series of techniques essentially based on target-shape imaging have been implemented. The target-shape image is actually a plot of range-spectrum reflectivity over a 2-D target projection. The proposed system features dual-dimension recognition of gesture components and high concurrency, applicable to edge computing scenarios. The satisfactory target-shape recognition rate optimized by image sequence joint analysis exhibits promising potential of real-time acoustic imaging in diverse application scenarios.

Acknowledgement

This research was supported by the National Key $R\&D$ Program No. 2020YFB1707700, the National Natural Science Foundation of China Grant No. 61902367, 62172377 and 41976185. This work is also dedicated to Chao Liu's newborn angel Shulang Liu.

References

[1] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *MobiCom*. ACM, 2016, pp. 82–94.

[2] S. Yun, Y. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *MobiSys*. ACM, 2017, pp. 15–28.

[3] R. Nandakumar, V. Iyer, D. S. Tan, and S. Gollakota, "Fingerio: Using active sonar for fine-grained finger tracking," in *CHI*. ACM, 2016, pp. 1515–1525.

[4] C. Liu, P. Wang, R. Jiang, and Y. Zhu, "AMT: Acoustic multi-target tracking with smartphone mimo system," in *INFOCOM*. IEEE, 2021.

[5] S. Yun, Y. Chen, and L. Qiu, "Turning a mobile device into a mouse in the air," in *MobiSys*. ACM, 2015, pp. 15–29.

[6] W. Mao, M. Wang, W. Sun, L. Qiu, S. Pradhan, and Y. Chen, "Rnn-based room scale hand motion tracking," in *MobiCom*. ACM, 2019, pp. 38:1–38:16.

[7] P. Cheng, I. E. Bagci, U. Roedig, and J. Yan, "Sonarsnoop: Active acoustic side-channel attacks," *CoRR*, vol. abs/1808.10250, 2018.

[8] X. Xu, J. Yu, Y. Chen, Y. Zhu, and M. Li, "Leveraging acoustic signals for vehicle steering tracking with smartphones," *IEEE Trans. Mob. Comput.*, vol. 19, no. 4, pp. 865–879, 2020.

[9] C. Cai, Z. Chen, H. Pu, L. Ye, M. Hu, and J. Luo, "AcuTe: Acoustic Thermometer Empowered by a Single Smartphone," in *SenSys*. ACM, 2020, pp. 91–104.

[10] 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 *UbiComp*. ACM, 2016, pp. 474–485.

[11] K. Ling, H. Dai, Y. Liu, and A. X. Liu, "Ultragesture: Fine-grained gesture sensing and recognition," in *SECON*. IEEE, 2018, pp. 28–36.

[12] Y. Wang, J. Shen, and Y. Zheng, "Push the limit of acoustic gesture recognition," in *INFOCOM*. IEEE, 2020, pp. 566–575.

[13] Y. Zhang, W.-H. Huang, C.-Y. Yang, W.-P. Wang, Y.-C. Chen, C.-W. You, D.-Y. Huang, G. Xue, and J. Yu, "Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, pp. 37:1–37:26, 2020.

[14] S. Gupta, D. Morris, S. N. Patel, and D. S. Tan, "Soundwave: using the doppler effect to sense gestures," in *CHI*. ACM, 2012, pp. 1911–1914.

[15] Y. Gao, Y. Jin, J. Li, S. Choi, and Z. Jin, "Echowhisper: Exploring an acoustic-based silent speech interface for smartphone users," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 3, pp. 80:1–80:27, 2020.

[16] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, L. Kong, and M. Li, "Lip reading-based user authentication through acoustic sensing on smartphones," *IEEE/ACM Trans. Netw.*, vol. 27, no. 1, pp. 447–460, 2019.

[17] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Battracker: High precision infrastructure-free mobile device tracking in indoor environments," in *SenSys*. ACM, 2017, pp. 13:1–13:14.

[18] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, "Beepbeep: a high accuracy acoustic ranging system using COTS mobile devices," in *SenSys*. ACM, 2007, pp. 1–14.

[19] H. Zhu, Y. Zhang, Z. Liu, S. Chang, and Y. Chen, "Hyperear: Indoor remote object finding with a single phone," in *ICDCS*. IEEE, 2019, pp. 678–687.

[20] K. Sun, T. Zhao, W. Wang, and L. Xie, "Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals," in *MobiCom*. ACM, 2018, pp. 591–605.

[21] D. Graham, G. Simmons, D. T. Nguyen, and G. Zhou, "A software-based sonar ranging sensor for smart phones," *IEEE Internet Things J.*, vol. 2, no. 6, pp. 479–489, 2015.

[22] H. Chen, F. Li, and Y. Wang, "Echotrack: Acoustic device-free hand tracking on smart phones," in *INFOCOM*. IEEE, 2017, pp. 1–9.

[23] H. Cheng and W. Lou, "Push the limit of device-free acoustic sensing on commercial mobile devices," in *INFOCOM*. IEEE, 2021.

[24] A. Zverev, "Digital mti radar filters," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 3, pp. 422–432, 1968.

[25] B. T. Kelley and V. K. Madisetti, "The fast discrete radon transform. i. theory," *IEEE Trans. Image Process.*, vol. 2, no. 3, pp. 382–400, 1993.

[26] S. Qian and D. Chen, "Joint time-frequency analysis: methods and applications," *Prentice-Hall, Inc.*, 1998.

[27] B. Boashash, "Note on the use of the wigner distribution for time-frequency signal analysis," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 36, no. 9, pp. 1518–1521, 2002.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, 2016, pp. 770–778.

[29] W. Mao, M. Wang, and L. Qiu, "Aim: Acoustic imaging on a mobile," in *MobiSys*. ACM, 2018, pp. 468–481.

[30] C. Zhang, F. Li, J. Luo, and Y. He, "ilocscan: Harnessing multipath for simultaneous indoor sourc e localization and space scanning," in *SenSys*. ACM, 2014, pp. 91–104.