Robust Human Face Authentication Leveraging Acoustic Sensing on Smartphones

Bing Zhou, Zongxing Xie, Yinuo Zhang, Jay Lohokare, Ruipeng Gao, Fan Ye

Abstract— User authentication on smartphones is the key to many applications, which must satisfy both security and convenience. We propose a novel user authentication system *EchoPrint*, which leverages acoustics and vision for secure and convenient user authentication, without requiring any special hardware. *EchoPrint* actively emits almost inaudible acoustic signals from the earpiece speaker to "illuminate" the user's face and authenticates the user by the unique features extracted from the echoes bouncing off the 3D facial contour. To combat changes in phone-holding poses thus echoes, a Convolutional Neural Network (CNN) is trained to extract reliable acoustic features, which are further combined with visual facial features extracted from state-of-the-art face recognition deep models to feed a binary Support Vector Machine (SVM) classifier for final authentication. Because the echo features depend on 3D facial geometries, *EchoPrint* is not easily spoofed by images or videos like 2D visual face recognition systems. It needs only commodity hardware, thus avoiding the extra costs of special sensors in solutions like FaceID. Experiments with 62 volunteers and non-human objects such as images, photos, and sculptures show that *EchoPrint* achieves 93.75% balanced accuracy and 93.50% F-score, while the average precision is 98.05% using acoustic features and basic facial landmarks. The precision is further improved to 99.96% with sophisticated visual features.

Index Terms—Mobile Sensing; Acoustics; Authentication; Face Recognition

1 Introduction

User authentication on smartphones is pivotal to many important daily apps, such as social networks, shopping and banking [2], [3]. Central to the user authentication is the balancing art between security and convenience. A solution must be secure while easy to use. A series of efforts have been undertaken to address this problem.

The most basic and traditional method, PIN number, has both usability (e.g., the pass code forgotten by the user) and security (e.g., shoulder-surfing [4] attacks) issues. Existing vision based approaches such as face recognition based authentication can be easily spoofed by images or videos of the user [5]. Simple twists such as requiring eye blinks are vulnerable to video attacks [6]. This is mainly caused by the lack of 3D information in images/videos of human faces. Iris scan [7] is probably the most secure way, however it requires special sensors unavailable on most mobile devices. Fingerprint sensors [8], [9], while convenient for authentication, are facing the challenge posed by the trend of ever-increasing screen size, which leaves little space for fingerprint sensors.

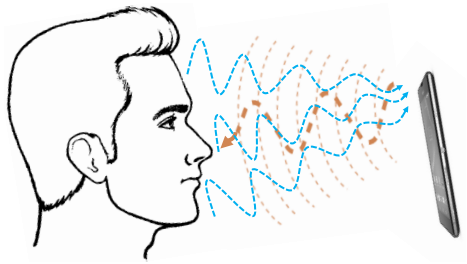


Fig. 1. *EchoPrint* emits nearly inaudible sound signals from the earpiece speaker to "illuminate" the user's face. The extracted acoustic features from the echoes are combined with visual features extracted from state-of-the-art face recognition models to authenticate the user.

The latest effort, Apple's FaceID [10], packs a dot projector, a flood illuminator and an infrared depth sensor in a small area to sense the 3D shape of the face, thus achieving high security while saving space. However, the special sensors still take precious frontal space and cost extra ($\sim 5\%$ of its bill of materials) [11]. Without such dedicated sensors for 3D sensing, facial recognition using 2D RGB images can not provide good accuracy and robustness for practical use. Thus, we ask this question: is an alternative using existing sensors possible?

In this paper, we propose a novel user authentication system *EchoPrint*, which leverages existing earpiece speaker and frontal camera thus can be readily deployed on most phones. It does not require costly special sensors (e.g., depth or iris) that take more spaces. *EchoPrint* combines acoustic features from a customized CNN feature extractor and visual features (basic facial landmarks and sophisticated visual features from deep neural networks) as the joint feature description of the user's face. It does not require the

- *A portion of this work was published in ACM MobiCom'18 [1] proceedings. This work was done when B. Zhou was with Stony Brook University.*
- *B. Zhou is with IBM Research, Yorktown Heights, NY, 10598, USA. E-mail: bing.zhou@ibm.com*
- *Z. Xie and F. Ye are with the ECE Department, Stony Brook University, Stony Brook, NY, 11790, USA. E-mail: {zongxing.xie, fan.ye}@stonybrook.edu*
- *Y. Zhang and J. Lohokare are with the CS Department, Stony Brook University, Stony Brook, NY, 11790, USA. E-mail: {yinuo.zhang, jay.lohokare}@stonybrook.edu*
- *R. Gao is with the School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China. E-mail: rpgao@bjtu.edu.cn*

user to remember any passcode, thus avoiding the usability issues as PIN numbers. The acoustic features depend on 3D facial geometries, thus it is resilient to image/video attacks easily spoofing 2D visual based approaches. Similar to FaceID, it does not require direct touch from the user, thus avoiding issues like wet fingers that pose difficulties to fingerprint sensors.

To achieve resilient, secure and easy-to-use authentication using acoustic and vision, we must address several challenges: i) echo signals are highly sensitive to the relative position between the user's face and the device (i.e., pose), which makes it extremely hard to extract reliable pose-insensitive features for robust authentication; ii) smartphones come with multiple speakers and microphones - which ones are most suitable, and what are the proper sound signals, are critical to authentication performance; iii) sophisticated signal processing, feature extraction and machine learning techniques are needed for fast user registration and real-time authentication.

We make the following contributions in this work:

- We design acoustic emitting signal suitable for considerations including hardware limitation, sensing resolution, and audibility to humans. We also create acoustic signal processing techniques for reliable segmentation of echoes from the face.
- We propose an end-to-end hybrid machine learning framework, which extracts representative acoustic features using a customized convolutional neural network, and fuses them with visual features extracted from state-of-the-art deep face recognition models to feed SVM for final authentication.
- We design a data augmentation scheme for generating "synthesized" training samples, which reduces false negatives significantly with limited training sample size, thus saving the user efforts in new profile registration.
- We build a prototype, conduct extensive experiments with 62 volunteers and non-human objects and find that *EchoPrint* achieves 93.75% balanced accuracy and 93.50% F-score, while the precision is up to 98.05%. The precision is further improved to 99.96% with sophisticated visual features. No image/video based attack is observed to succeed in spoofing our system.

To the best of our knowledge, *EchoPrint* is the first to leverage active acoustic sensing combined with vision features for smartphone user authentication, demonstrating robust performance without requiring any additional special sensor.

2 Background

2.1 Attack Model

We summarize typical attack scenarios for major existing authentication methods. *1) Replay Attack.* 2D image based face recognition systems suffer from replay attacks by images or videos of the user face. The face recognition system on Samsung's flagship Galaxy S8 is reported to be spoofed by a simple picture [12]. *2) Shoulder-surfing Attack.* When the victim user performs PIN number authentication, it can be easily exposed to shoulder-surfing attacks [4], which has been a common case for someone standing close by to peek the whole PIN typing. *3) Biometric Duplication Attack.* Fingerprint is the mainstream biometric used for authentication solutions. However, fingerprints are widely left on objects (e.g., glasses) touched by the user, and can be duplicated with reasonable efforts and skill [8] to fool the sensor.

We assume that the adversary has no prior knowledge about the victim's authentication information, and it can only be captured during the victim user's operation. The proposed face authentication method, being an alternative to the traditional PIN number and fingerprint authentication methods, prevents the victim user from the shoulder-surfing attack on the PIN number and biometric duplication attack on the fingerprint. Technically, the adversary can spoof the authentication system with *Biometric Duplication Attack* on facial feature. However, it becomes prohibitively intractable with a strict requirement on the combination of geometry, materials, and fabrication such that a facial sculpture can faithfully mimic the acoustic reflection features of the authenticated user. The *Replay Attack* is the most concerned adversarial model, as it could be possible to capture the victim's image and acoustic data during her/his authentication operation. We present more discussions and experiment results regarding the attack models in our evaluation section 8.5.

2.2 Design Considerations

We make the following considerations when designing *EchoPrint*. *1) Universal.* We prefer to use existing hardware widely available on most smartphones, so that it can be deployed at large scale rapidly with minimum hardware costs. Besides, we want to use a biometric that is pervasive to every human being. *2) Unique.* The human face has been widely used as a biometric because it is distinctive. However, most existing 2D visual based systems can be spoofed by images or videos. Thus we leverage the 3D information of the facial contour for much higher security. *3) Persistent.* The biometric must not change much over time. Biometrics such as heart beat, breathing, gait are highly affected by the user's physical conditions (e.g., running vs. walking), thus not optimal choices for robust authentication. In contrast, the human face geometries are not likely to change too much over short time periods. However, daily changes like wearing hats or glasses must be easily accommodated. *4) Difficult to Circumvent.* This is essential for any authentication system to ensure a high security level. Existing authentication approaches, such as PIN numbers, 2D based face recognition, fingerprint sensors still have quite some risks to be circumvented. Because our two-factor authentication examines both acoustic and visual features simultaneously, circumventing would require duplicating both 3D facial geometries and acoustic reflections properties close enough to the human face, which will be much more difficult than needed in circumventing other methods.

2.3 Design Goal

Based on the above discussions, our goal is to build a highly secure, resilient two-factor authentication system

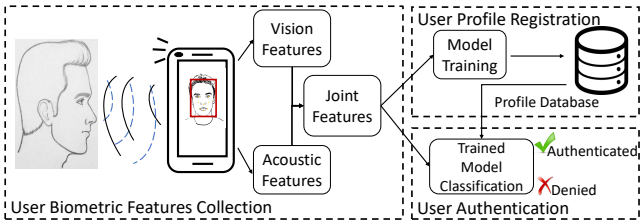


Fig. 2. The smartphone actively emits sound signal towards the user's face, and collects image and echo data for authentication. Sophisticated visual features from face recognition models and acoustic features extracted from a customized CNN are jointly used for the final classification.

that is available to most existing smartphones without requiring any special sensors. To suit different application scenarios, we design different authentication modes that have proper trade-offs between security, convenience and power consumption. While light weight vision algorithms should be used to minimize computation complexity, we keep the option open for future integration with state-of-the-art vision solutions. We believe such "free" acoustic-aided authentication will play an important role in mobile authentication developments.

3 Overview

EchoPrint uses speakers/microphones for acoustic sensing and the frontal camera for facial landmarks detection. It extracts acoustic features from echo signals using a deep learning approach and fuses such features with visual features extracted from state-of-the-art face recognition models as a joint representation for authentication. Figure 2 shows the overview of the system design, which consists of two major phases: user *registration* and user *authentication*.

In the *registration phase*, *EchoPrint* detects facial landmarks (e.g., eyes, mouth) using the frontal camera. Meanwhile, the earpiece speaker emits designed acoustic signals to "illuminate" the user's face. Echoes bouncing back are received by the microphone. A pre-trained CNN model is used to extract acoustic features resilient to phone pose changes, which are combined with visual features as joint feature representation, and fed into an SVM classifier for model training. In the *authentication phase*, the user just needs to hold the smartphone in front of the face for facial landmarks detection and acoustic sensing. The joint features are extracted and fed into the trained SVM classifier for final authentication.

4 Acoustic Sensing

Acoustic echoes from the human face are highly distinctive: i) the echoes are very sensitive to the relative position between the user face and device. ii) each 3D facial contour is a unique set of multiple reflecting surfaces [13], which create a unique sum of individual echoes. iii) different materials absorb, attenuate sound waves differently, allowing us to distinguish objects of similar geometry but different materials (e.g., a stone sculpture).

4.1 Speaker/Microphone Selection

There are two speakers, a main one at the bottom and an earpiece speaker at the top for hearing phone calls. There are also one microphone at the bottom, and another at the top for noise cancellation [14]. We select the earpiece speaker, top microphone, and frontal camera combination for robust acoustic/visual sensing. The earpiece speaker is chosen for sound emitting for two reasons: i) it's a highly standard design on almost all existing smartphones. Its location is suitable for "illuminating" the user's face; whereas the main speaker has a more diverse design, either located at the bottom or on the back; ii) the earpiece speaker is close to the frontal camera, which minimizes alignment errors when the frontal camera is used for adjusting the phone pose. The top microphone is chosen as the receiver because it is close to the earpiece speaker, and it's less affected by the user's hand holding the device.

4.2 Acoustic Signal Design

There are several considerations in the emitting signal design. First, it should facilitate isolation of the segment of interest (i.e., echoes from the face) from the other reflections, such as interferences from clutters and self-interference from the speaker. This requires the signal be short enough so that echoes from objects at different distances have little overlap in time domain. Second, the acoustic signal should be as inaudible as possible to human ears to minimize annoyance. An ideal frequency range should be over $20KHz$. Lastly, the designed signal frequency range should be apart from ambient noises (usually under $8KHz$), to enable noise removal (e.g., using band-pass filters) and improve robustness.

According to our survey, a comfortable distance from human eyes to the phone is 25 - $50cm$, corresponding to a time delay of $\sim$1.4 - $2.8ms$ at the speed of sound. From our experiments, when the frequency goes above $20KHz$, serious power attenuation and worse signal to noise ratio occur, thus echoes from faces are buried under noises. Considering all these facts, we choose a pulse signal with a length of $1ms$ with linear increasing frequencies from 16 - $22KHz$. A Hanning window [15] is applied to reshape the pulse envelop to increase its peak to side lobe ratio, thus producing higher SNR for echoes. For authentication modes that require continuous sound emitting, we leave a delay of $50ms$ for each pulse such that echoes from two consecutive pulses do not overlap.

4.3 Acoustic Signal Pre-processing

4.3.1 Background Noise Removal

The received raw signal goes through a 16 - $22KHz$ Butterworth band-pass filter to remove background noises, such that weak echoes from human faces will not be buried in the noise.

A sample recording segment of a received signal after noise removal is shown in Figure 3. The *direct path* segment is the emitting signal traveling from speaker to the microphone directly, which ideally should be a copy of the emitting signal and has the highest amplitude. The *major echo* corresponds to the mix of echoes from the major surfaces (e.g., cheek, forehead) of the face. Other surfaces of the

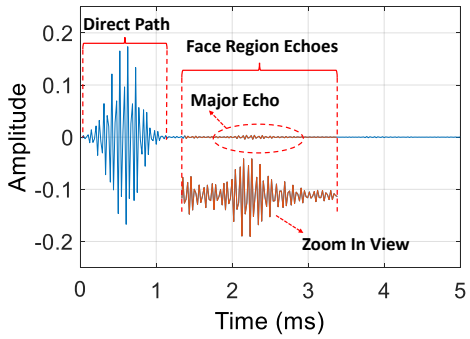


Fig. 3. Sample recording segment of a received signal after noise removal.

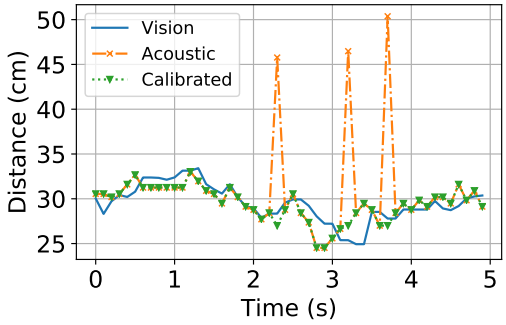


Fig. 4. Distance measurements from acoustics, vision, and calibrated acoustics.

face (e.g., nose, chin) at different distances to the phone also produce echoes, arriving earlier/later than the major echo. The *face region echoes* include all these echoes, capturing the full information of the face. Accurate segmenting the face region echoes is critical to minimize the disturbances from dynamic clutters around the phone, and reduce the data dimension for model training and performance.

4.3.2 Signal Segmentation

There are two steps extracting the face region segment: locating the direct path segment in raw recordings, then locating the major echo thus face region segment after the direct path segment.

Locating the Direct Path. An easy but naive assumption is that a constant gap exists between the emitting and recording, thus the direct path can be located after that constant gap. However, both emitting and recording must go through multiple layers of hardware and software processing in the OS, many of which have unpredictable, varying delays. Thus locating the direct path using a constant delay is extremely unreliable. Instead, since the direct path signal usually has the highest amplitude, using cross-correlation to locate it is more reliable [16]. From our experiments, occasional offsets of direct path signal still happen after cross-correlation, due to ambiguities from comparable peak values in the cross-correlation result. We propose two techniques to enhance the stability:

i) Template Signal Calibration. Due to the hardware (speaker/ microphone) imperfection, the received sound signal is usually slightly different from the designed emitting signal. To get an accurate "template" signal for cross-correlation, we perform emitting and recording in a quiet environment, so that the direct path signal can be reliably detected and saved as a calibrated template for future cross-correlation.

ii) Signal Fine-tuning. In addition to the Hanning window, we manually tune the signal slightly to make the key peaks/valleys more prominent, which reduces cross-correlation ambiguity significantly. Only the central portion (15 samples) of the template signal is used in cross-correlation, further enhancing resilience to residual noises.

Locating the Major Echo. A straightforward way for locating the major echo is to find cross-correlation peak location corresponding to typical phone holding distance (e.g., 25 - 50cm) after the direct path location. However, human face echoes can be so weak that echoes from larger obstacles faraway can have comparable amplitudes. This makes the estimation unstable and leads to occasional location "jumping", thus outliers in distance measurements. The dotted line in Figure 4 shows the distance measurements from acoustic while the device is moving back and forth from the face. We can observe quite some outliers due to such "jumping" outliers. To solve this problem, we propose a vision-aided major echo locating technique of two steps:

i) Vision Measurement Calibration. From camera image projection principle, the closer the device to the face, the larger the image and larger distances between facial landmarks, and vice versa. Thus the distance from face to device d_v can be formulated as $d_v = \tau \cdot \frac{1}{d_p}$, where d_p is the distance between two facial landmarks and τ is an unknown scale factor specific to the user. We choose d_p as the pixel distance between two eye landmarks as they are widely separated and can be detected reliably. To estimate the scale factor τ, we calculate τ_i for each pair-wise $d'_{v,i}$ from acoustic distance measurement and $d_{p,i}$ in pixels. To eliminate errors caused by acoustic distance measurement outliers, we first find the major cluster of $\{\tau_i\}$ using density-based spatial clustering algorithm DBSCAN [17], then leverage linear regression to find the best τ that minimizes the offset between d'_v and $\tau \cdot \frac{1}{d_p}$. Figure 4 shows that outliers are removed in vision calibrated acoustic distance measurements.

ii) Vision-aided Major Echo Locating. Although vision based distance measurement is more stable than acoustics, it can not capture the error caused by rotations of smartphone or user's face. Thus the vision calibrated distance measurement is used to narrow down the major echo searching range and reduce outliers. We still use cross-correlation to find the exact major peak location within this range. Note that the user face cannot rotate to extreme angles, otherwise facial landmark detection may fail.

Face Region Echoes. Since the depth of human face is limited, we extend 10 sample points before and after the major echo segment to cover the whole face region (allowing a depth range of $\sim 7cm$), which are later used as inputs for machine models for authentication.

4.4 Segmented Signal Analysis

The face region echoes are a combination of individual echoes with different amplitudes and phases, thus isolating individual echoes in time domain can be very hard due to noises. Instead, we measure the arrival time of

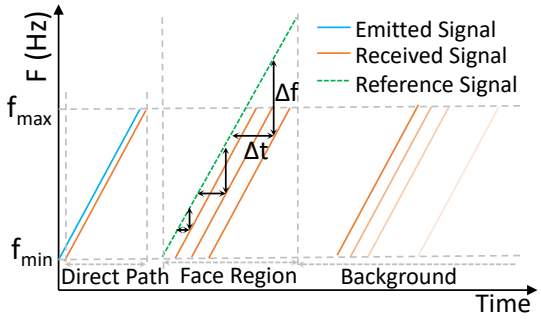


Fig. 5. Illustration of FMCW.

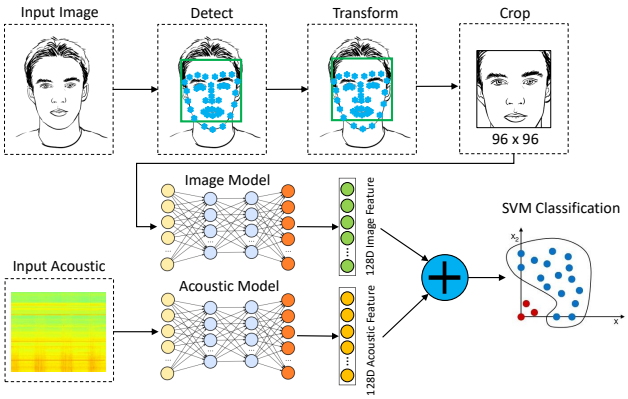


Fig. 6. The authentication framework consists of three major components: image model, acoustic model and SVM classifier.

each echo by a technique Frequency-Modulated Continuous Wave (FMCW) [18] used in radars. In traditional FMCW, the speaker transmits continuous chirp signals with linear increasing frequency, from f_{min} to f_{max}. To estimate the distance from an object, FMCW compares the frequency of the echo signal to that of a reference signal using a technique called signal mixing, to find the frequency shift Δf (shown in Figure 5), which is proportional to the distance. Thus finding Δf gives the distance (i.e., Δf multiplying a constant coefficient).

To capture minute surface geometries on the face, the FMCW distance measurement resolution is critical. The resolution in Δf is equal to the size of one bin in the FFT, which depends on the bandwidth used. This is why we use a wide frequency of 16 - 22KHz, though it may be slightly audible to some users. In Figure 5, the FFT is taken over a duration of the face region with length T and hence the size of one FFT bin is 1/T. Given a minimum measurable frequency shift $\Delta f_{min} = 1/T$, the minimum measurable distance resolution can be computed using the slope of signals (see Figure 5), which is the total swept bandwidth B divided by the sweep time T. Thus the distance resolution:

$$d_r = C\frac{TOF_{min}}{2} = C\frac{\Delta f_{min}}{2 \times slope} = \frac{C}{2B} \quad (1)$$

where C is the speed of sound. Assuming $C = 343m/s$ at 20° Celsius, thus d_r is $\frac{343m/s}{2\times6000s^{-1}} = 2.88cm$. Note that this is the resolution that FMCW can separate mixed echoes. The resolution of major echo location corresponds to one single acoustic sample, which is $\frac{C}{2F_s} = 3.57mm$, where $F_s = 48KHz$ is the recording sampling frequency. The spectrogram of the segmented face region echoes after FMCW signal mixing is then used as input for CNN training in Section 5.1.

5 Acoustic and Visual Feature Extraction

We design an end-to-end hybrid machine learning framework for authentication, which consists of three major components (shown in Figure 6): a CNN based acoustic feature extraction model, a visual feature extraction model, and an SVM classifier for two-factor authentication.

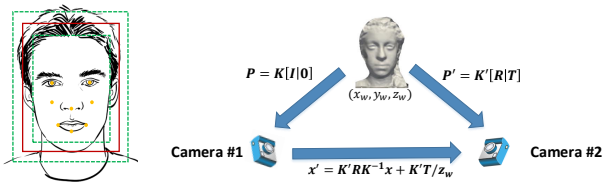
5.1 Acoustic Feature Extraction

Traditional acoustic features such as mel-frequency cepstral coefficients [19], chromagram [20] and spectral contrast [21] have been proven to be effective in human speech recognition and voice-based authentication, but not in active acoustic sensing as in our case. Recently, deep learning approaches (especially CNNs) have shown great successes in a variety of challenging tasks such as image classification due to their powerful automatic feature extraction [22], [23]. We design a CNN based neural network which takes the spectrogram of the segmented signal as input, and train it on a large data set collected from users. We find that such extracted features outperform all traditional features (in Section 8.2).

A customized CNN architecture designed for acoustic feature learning is shown in Table 1. We use rectified linear unit (ReLU) as activation function for convolutional layers, a popular choice especially for deep networks to speed up training. Two max pooling layers with a size of 2×2 are used to down-sample the input representations from their previous activation layers. This saves computational costs by reducing the number of parameters for both training and inference, which is critical when the model needs to be deployed on mobile devices. Dropout layers are added after each max pooling layer to prevent over-fitting. Batch normalization normalizes the output of a previous layer by subtracting the batch mean and dividing by the batch standard deviation, which increase the stability of the neural network and speed up training ($\sim 6\times$ speedup in our case). Categorical cross-entropy is used as the loss function. The dense layer with softmax activation function outputs the probability of each class. The CNN is trained on a data set that contains acoustic samples from 50 classes (45 users and 5 non-human classes). Note that although the CNN is trained for 50 classes, the objective of the trained model is to extract features that can be used to distinguish far more classes beyond those 50. To use the trained model as a general acoustic feature extractor, the last layer, which is used for final classification, is removed. Thus the remaining network outputs a 128 dimensional feature vector. The trained model has 710539 parameters, and a size of $5.47MB$, which is portable enough for mobile devices for real-time inference.

TABLE 1
CNN layers and parameter amounts.

Layer	Layer Type	Output Shape	# Param
1	Conv2D + ReLU	(33,61,32)	320
2	Conv2D + ReLU	(31,59,32)	9248
3	Max Pooling	(15,29,32)	
4	Dropout	(15,29,32)	
5	Batch Normalization	(15,29,32)	128
6	Conv2D + ReLU	(15,29,64)	18496
7	Conv2D + ReLU	(13,27,64)	36928
8	Max Pooling	(6,13,64)	
9	Dropout	(6,13,64)	
10	Batch Normalization	(6,13,64)	256
11	Flatten	(4992)	
12	Dense + ReLU	(128)	639104
13	Batch Normalization	(128)	512
14	Dense + Softmax	(50)	5547



(a) Landmark detection. (b) Landmark transformation.

Fig. 7. Facial landmarks and face tracking, and landmark transformation between two camera positions.

5.2 Visual Feature Extraction

5.2.1 Basic Facial Landmarks Detection

We extract lightweight visual features of the face to complement acoustic ones. The vision techniques serve two purposes: i) we detect facial landmarks which are later used as basic visual features. ii) we track the user's face on the smartphone screen so that the user can hold the device within some "valid" zone (thus distance and orientation) for data collection.

We detect the 2D coordinates of facial landmarks (e.g., corners/tips of eyes, nose and mouth) on the image as features, using the mobile vision API from Google [24] on Android platform (shown in Figure 7(a)). The face is tracked by a bounding rectangle. We observe that these landmarks describe critical geometry features on the face, and their locations are associated with the relative position from the device to the face. Detailed description of the implementation is presented in Section 7.

5.2.2 Sophisticated Visual Features Extraction

Sophisticated visual features from state-of-the-art face recognition models such as OpenFace [25] can provide more rich and robust information from the visual aspect. As we use the images for facial key points extraction, we propose a method to further extract more sophisticated visual features to enhance the overall security of our authentication. Getting a low-dimensional visual representation is crucial for efficient classification on mobile devices where the resource is limited. As the intrapersonal image variations such as angles, distances and even facial expressions can cause difficulty in classification, we adjust and normalize the face before the actual feature extraction.

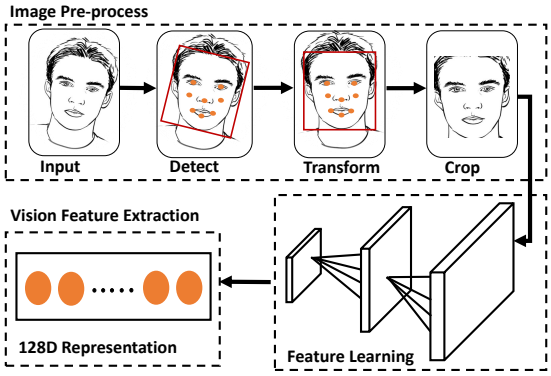


Fig. 8. Image pre-processing and visual feature extraction.

Image Pre-processing. Figure 8 shows the four stages to pre-process the image input for training the face representation neural network. On the mobile device, we get the aligned face image input once the App detects there appears the face in the red box while the red box stays in between the green boxes as shown in Figure 9. We can drastically reduce the computation effort for further image processing given alignment by the first stage. In the second stage, we aim to locate where the eyes, nose and lips are. To mitigate the constraints from illumination conditions, we leverage a pre-trained detector based on Histogram of Oriented Gradients to layout the face landmarks. To make it easier for facial recognition, thus authentication, we project all the face landmarks to our predefined positions using affine transformations, which have the expression as shown in the Equation 2:

$$T = A_{2\times2} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + B_{2\times1} = M_{2\times3} \cdot [x, y, 1]^T \qquad (2)$$

$$M_{2\times3} = \begin{bmatrix} A_{2\times2} & B_{2\times1} \end{bmatrix}$$

where M is obtained from the relationship between predefined landmark positions and detected positions of landmarks in raw image inputs, and it conducts rotation, translation and scale operations [26]. Thus no matter in what angle the sample is taken, we can achieve roughly the same positions for every face landmark. The affine transformation provides an affine mapping from the raw image to a well frontalized and normalized image input for training. Obtaining the image with landmarks of known positions, we crop the picture to a more compact image thus further reducing the complexity in training.

Image Feature Extraction. Now that we achieve the smaller size of the normalized input space, which is suitable as the input to the deep convolutional network to achieve a desirable low-dimensional representation, which can generalize well to faces that are new to the neural network. We achieve this goal by taking advantage of OpenFace's neural network [25], which is a reduced version of *nn4* proposed by Google's FaceNet [27]. The network is trained by using a combination of classification and *triplet loss*, which minimizes the distance between faces of the same identity and enforces a margin between the different identities. After completion of training, we leverage the pre-trained model

as a feature extractor to map the face image input to a 128-dimensional embedding space, in which faces from the same identity should be close together and form well separated clusters, such that they can be easily recognized/classified. And the extracted visual features later will be combined with acoustic features to form a joint embedding, which is a generic representation for anybody's face, for classification (i.e., final authentication).

5.3 Two-factor Authentication

The joint acoustic and visual features are used for two-factor authentication.

Features Summary. To use basic visual features, the facial landmarks on the image are concatenated with the corresponding 128-dimensional CNN features as the joint features for final authentication. Both acoustic and vision data are collected simultaneously so that they are well synchronized, which ensures the correspondence between facial landmarks distribution on the screen and relative device position, thus echo signals. As an option, we also concatenate the 128-dimensional sophisticated visual feature and acoustic feature for even higher performance at a cost of more computational complexity.

Classifier. One-class SVM is an unsupervised algorithm that learns a decision function for novelty detection: classifying new data as similar or different to the training set [28]. It detects the soft boundary of the training set so as to classify new samples as belonging to that set or not. We use one-class SVM with radial basis function (RBF) kernel function for final classification. It allows us to train an SVM classifying model for a new user (or the same user wearing new hats or glasses) on mobile devices easily, without requiring large amounts of training data as in CNN.

Ideally, a user should move the device at various relative positions to the face so as to collect sufficient training data during user registration. In practice, this imposes more efforts on the user, and it is hard to tell when sufficient data has been collected. Insufficient training data will cause higher false negatives (i.e., denial of the legitimate user). Thus we propose a data augmentation technique, which populates the training data by generating "synthesized" training samples based on facial landmark transformation and acoustic signal prediction. During this process, we transform measured facial landmarks and acoustic signals into synthesized ones, by assuming different poses of the phone.

5.4 Data Augmentation

Data augmentation is commonly used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data [29]. It is an effective way to prevent overfitting when the data amount is relatively small. In our design, we propose a data augmentation technique based on camera projective geometry and sound propagation inverse-square law.

In projective geometry, the projection matrix $\boldsymbol{P}$ of a 3D point (x_w, y_w, z_w) in the world coordinate system onto the image plane in camera is modeled as Equation 3:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = \boldsymbol{K} \cdot [\boldsymbol{R}|\boldsymbol{T}] \cdot \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = \boldsymbol{P} \cdot \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{3}$$

where λ is the scale factor for homogeneous coordinates, (u, v) denotes its pixel coordinate on image, $\boldsymbol{K} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$ is the intrinsic matrix of the camera, e.g., the focal length f_x and f_y, skew s, and image center (c_x, c_y) in pixels. $[\boldsymbol{R}|\boldsymbol{T}]$ represents the extrinsic matrix of the camera, i.e., camera's pose in the world coordinate system, where $\boldsymbol{R}$ is a 3×3 matrix for its 3D orientation, and $\boldsymbol{T}$ is a 3×1 matrix for its 3D translation.

As shown in Figure 7(b), assume two cameras take images of the same object at different distances/angles, and $\boldsymbol{x} = [u, v, 1]^T$ and $\boldsymbol{x'} = [u', v', 1]^T$ represent the object's pixel coordinates on two images. Without loss of generality, we define the first camera as the world origin, thus the projection matrix of two cameras are:

$$\boldsymbol{P} = \boldsymbol{K}[\boldsymbol{I}|\boldsymbol{0}], \boldsymbol{P'} = \boldsymbol{K'}[\boldsymbol{R}|\boldsymbol{T}] \tag{4}$$

where $\boldsymbol{I}$ is a 3×3 identity matrix.

Based on the above background of projective geometry, we can transform the landmark pixel coordinates in one camera to those of any new camera pose, thus augmenting our training set.

Step 1: Compute the landmark's world coordinates. Given the projection matrix $\boldsymbol{P}$ and landmark pixel coordinates $\boldsymbol{x}$ of the first camera, we can compute the landmark's world coordinates as $(x_w, y_w, z_w)^T = z_w \boldsymbol{K}^{-1} \boldsymbol{x}$, where z_w is the distance of landmark from camera center, which can be measured via our acoustic sensing.

Step 2: Transform the landmark onto new images. From the projection matrix of the new camera pose, we can compute the corresponding pixel coordinates of the landmark as:

$$\boldsymbol{x'} = \boldsymbol{K'}\boldsymbol{R}\boldsymbol{K}^{-1}\boldsymbol{x} + \boldsymbol{K'}\boldsymbol{T}/z_w \tag{5}$$

This transform equation consists of two parts: the first term depends on the image position alone, i.e., $\boldsymbol{x}$, but not the landmark's depth z_w; the second term depends on the depth and takes account of the camera translation. In the case of pure translation ($\boldsymbol{R} = \boldsymbol{I}$, $\boldsymbol{K'} = \boldsymbol{K}$), Equation 5 reduces to $\boldsymbol{x'} = \boldsymbol{x} + \boldsymbol{K}\boldsymbol{T}/z_w$.

Step 3: Data augmentation. We augment our training set based on Equation 5. i) Before data collection, we calibrate the camera with a benchmark paper printing of a chessboard with known size, thus obtain its intrinsic matrix $\boldsymbol{K}$. ii) Given a new camera pose of $\boldsymbol{\theta} = (\boldsymbol{T}, \boldsymbol{\phi})$, where $\boldsymbol{T}$ for its 3D coordinates and $\boldsymbol{\phi} = (\alpha, \beta, \gamma)$ for its rotation angles along three axes of the smartphone, we transform $\boldsymbol{\phi}$ to the 3×3

rotation matrix $\boldsymbol{R}$ based on Rodrigues's Formula [30], then compute $\boldsymbol{x}'$ via Equation 5.

Accordingly, following the sound propagation inverse-square law, the face region signal segment is shifted by the same distances, with the amplitude adjusted by the scale equal to the inverse of the square of distance. Due to the omni-directional property of smartphone speaker and microphone, a slight device rotation at a fixed position causes negligible changes in the signal, thus only device position change accounts for acoustic signal transform.

6 Authentication Modes

We propose three authentication modes: two-factor one-pass authentication, low-power continuous authentication and ultra low-power presence detection, suitable for scenarios requiring progressively less security level but more user convenience and power efficiency.

Two-factor One-pass Authentication. In this mode, the user must hold the phone properly to align his face within the valid area rectangle as shown on the screen (see Figure 9). Both visual facial landmarks from camera images and acoustic features extracted by the trained CNN are fed to the SVM for recognition. This incurs the most computation, energy costs, providing the highest security level suitable for scenarios such as phone unlock, account log in. In this mode, we can also leverage sophisticated visual features for higher performance. Due to the heavy computation of sophisticated visual feature extraction, images are offloaded to a server for inference computation.

Low-power Continuous Authentication (LP mode). In this mode, acoustic features extracted from the CNN are used in one-class SVM classification. It avoids power hungry cameras and real-time video processing, providing reduced security level suitable for scenarios such as continuous access/browse of private data in banking transactions after login is completed. The user needs to hold the phone in position ranges similar to training data collection.

Ultra Low-power Presence Detection (ULP mode). This mode uses acoustic signal only and an SVM model to detect the presence of the user face. To minimize computation and energy costs, the spectrum of a set of samples (e.g., the first 80 after the direct path signal) instead of CNN extracted features is fed to SVM. Data collected to train the SVM include positive samples while holding the device before the user's face, negative samples when putting the device on tables, in pockets, or holding it away from the user. This mode consumes the least power and is suitable for scenarios like auto screen lockup when the user face is not present.

7 Implementation

We implement *EchoPrint* on multiple Android smartphones, including SamSung S7 Edge, SamSung S8, and HuaWei P9. Figure 9 shows the UI. The prototype consists of three major modules: facial landmark detection, acoustic sensing, and machine learning pipeline for authentication.

Facial Landmark Detection and Feature Extaction. We use Google mobile vision API [24] for real-time facial landmark detection and face tracking. The middle red rectangle (Figure 9) denotes the detected face area, and two green

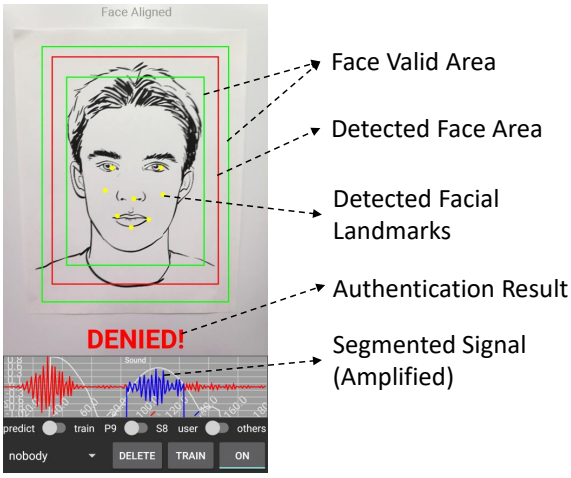


Fig. 9. *EchoPrint* prototype application UI.

rectangles are the predefined inner and outer bounds of facial valid areas. The user face must be aligned within the two green rectangles during data collection; otherwise the acoustic data are discarded. Yellow dots are detected facial landmarks, saved in pixel coordinates. For sophisticated visual feature extraction, we offload video frames to a server for inference computation.

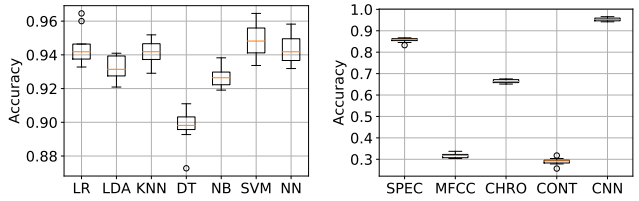
Acoustic Sensing. The acoustic signal is pre-processed and displayed on the screen in real-time, and the segmented signal from the face is highlighted in blue. For better visualization, we amplify the signal by $3\times$ after the direct path signal.

Machine Learning Pipeline. The machine learning pipeline requires one CNN acoustic feature extractor, one visual feature extractor, and one SVM classifier. We train the CNN model off-line on a PC with Intel i7-8700K CPU, 64GB memory and GTX 1080 Ti GPU. Keras [31] with Tensorflow [32] backend is used for CNN construction and training. The trained model is frozen and deployed on mobile devices. We adopt one of the state-of-the-art face recognition model – OpenFace [25] – as our visual feature extractor, which runs on the server for feature extraction. Using both acoustic and visual features extracted from deep neural networks, an SVM classifier using LibSVM [33] is trained on mobile devices. Both acoustic CNN and SVM inferences are performed on mobile devices in real-time.

8 Evaluation

8.1 Data Collection

We obtain the required human subjects training certificate from our institution before data collection. 45 participants of different ages, genders, and skin colors are recruited in experiments. The diversity in physical appearances of participant faces help us capture sufficient data to create a strong feature extraction model. We also include 5 non-human classes: printed/displayed human faces on different materials such as paper, desktop monitor, photo on paper box, wall and a marble sculpture. During data collection, each participant is asked to hold the smartphone in front of his/her face to ensure face alignment. To accommodate



(a) Different classifiers. (b) Different features.

Fig. 10. Different classifiers performance on extracted features from CNN, and SVM performance using different features.

slight phone movements, participants are encouraged to move the phone slowly to cover different poses. Data captured while the face is out of the valid area are discarded automatically. About 120 seconds' data is captured from each user, at around 7 - 8MB and containing $\sim$ 2000 samples. To ensure diversity, the data is collected in multiple uncontrolled environments (e.g., quiet laboratories, noisy classrooms, and outdoor environments) under different background noises and lighting conditions. A portion of the participants who are more accessible to us collected data in multiple sessions at different times and locations. Facial landmarks are also detected and recorded simultaneously, but no facial images are recorded to protect the participants' privacy. In total, the data set contains 91708 valid samples from 50 classes. We divide it in three parts, 70% for model training, 15% each for model validation and testing. Additionally, 12 more volunteers join as new users for model evaluation.

8.2 CNN Feature Extractor Performance

We compare the performance of different classifiers and feature extraction methods using the test data set.

Different Classifiers. The last fully connected layer of our trained CNN is removed so that the remaining network is used as a general feature extractor. Such extracted features are then fed to different classifiers for final classification. We compare Linear Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), Decision Tree (DT), Naive Bayesian (NB), Support Vector Machine (SVM) and a standalone Neural Network (NN). The box plot in Figure 10(a) shows the lower and upper quartiles, and the median. The whiskers extend from the box show the range of accuracy, and outliers beyond the whiskers are marked as circles. We find SVM outperforms all other classifiers, and it takes short time (15.06s compared to 65.38s of NN which has second best performance) for training. Thus we use SVM as the final classifier for authentication.

Different Features. We compare different commonly used acoustic features: spectrogram (SPEC), mel-frequency cepstral coefficients (MFCC) [19], chromagram (CHRO) [20], spectral contrast (CONT) [21] and our CNN features. Figure 10(b) shows their accuracies using SVM classifier. Our CNN extractor outperforms all other features and achieves the highest accuracy of $\sim$ 95%, which show the effectiveness and necessity of the CNN feature extractor. Spectrogram has less accuracy at $\sim$ 85%, and chromagram 67%. MFCC and CONT have much lower accuracy $\sim$ 30%, which is what we expected because they are mostly used for human voice recognition, not active acoustic sensing used in *EchoPrint*. Besides, the 15.06s using CNN features to train the SVM model is a fraction of the 134s needed when training with spectrogram. This is a significant improvement when training a model on resource-constraint mobile devices, which is critical for the speed of user registration.

8.3 Authentication Accuracy

In a binary classification problem, there are four results: true positive (TP), positive samples correctly classified as positive class; true negative (TN), negative samples correctly classified as negative class; false positive (FP), negative samples wrongly classified as positive class; false negative (FN), positive sample wrongly classified as negative class. Specifically, in authentication scenarios, a high TP means the authorized user can get access easily, a high TN means the system can block most attacks. The worst case is high FP, which means unauthorized users gain access. A high FN means the authorized user may be denied access, which is annoying and not user-friendly. In this evaluation, we train a one-class SVM for each subject and attack the model using the data from the rest users. Note that the model is trained on positive samples only, it does not have negative samples from attackers during training.

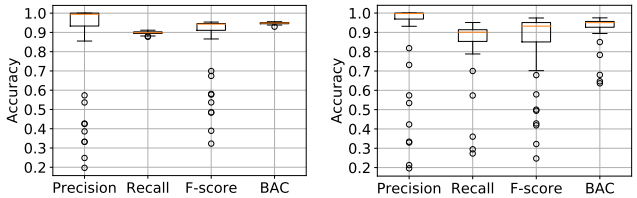
8.3.1 Precision, Recall, F-score and BAC

We introduce precision, recall, F-score and balanced accuracy (BAC) as metrics. Precision is the fraction of true positives among all samples classified as positive, defined as $P = \frac{TP}{TP+FP}$; recall is the fraction of true positives among all positive samples, defined as $R = \frac{TP}{TP+FN}$. A high precision means the authorized user can pass easily, a high recall means the authorized user is seldom denied. When the class distribution is imbalanced, precision and recall alone can be misleading. We also introduce F-score and balanced accuracy (BAC), both insensitive to class distribution. F-core is the harmonic mean of precision and recall with a best value of 1 and worst value of 0, defined as $F\text{-}score = 2\frac{P \cdot R}{P+R}$. BAC is the average of true positive rate ($TPR = \frac{TP}{TP+FN}$) and true negative rate ($TNR = \frac{TN}{TN+FP}$), defined as $BAC = \frac{1}{2} \cdot (TPR + TNR)$. A BAC of 1 means no false positive (i.e., successful attack) or false negative (i.e., denied access of legitimate users).

TABLE 2
Mean/median accuracy with vision, acoustic and joint features.

	Vision	Acoustic	Joint
Precision (%)	72.53 / 80.32	86.06 / 99.41	88.19 / 99.75
Recall (%)	64.05 / 64.04	89.82 / 89.84	84.08 / 90.10
F-score (%)	65.17 / 69.19	85.39 / 94.31	83.74 / 93.23
BAC (%)	81.78 / 81.83	94.79 / 94.88	91.92 / 95.04

Table 2 shows the mean and median accuracies using vision, acoustic, and joint features. Vision (2D coordinates of a few facial landmarks like the corners/tips of eyes, nose and mouth) is the worst with a low average precision of $\sim$ 72%. Acoustic achieves 86%, and joint features further increase it to 88% while also decreasing recall by $\sim$ 6%. That is because simple 2D coordinates of facial features do not capture the full characteristics of the face, thus alone they



(a) Acoustic features only. (b) Joint features.

Fig. 11. The precision, recall, F-score and BAC of one-class SVM model using acoustic and joint features.

do not perform well when many test subjects exist. They can help "block" unauthorized users which happen to have similar acoustic features, thus increasing precision. But they also make it harder for the authorized user to pass, thus decreasing recall. Both acoustic and joint features have an average F-score $\sim 85\%$ and BAC above 90%. The vision features used are not sophisticated and detailed visual features (e.g., the contour of face) of facial appearances as used in state-of-the-art vision-based face recognition systems. These basic face landmarks are mainly used for face alignment, which is critical for robust acoustic sensing. While such facial landmarks are not intended to greatly improve recognition accuracy, *EchoPrint* as an acoustic based approach is free to incorporate more sophisticated facial features, e.g., features from a deep neural network trained on a huge face image dataset [25]. Those would have a much higher impact on performance improvements.

Note that the median precision ($\sim 99\%$) and F-score ($\sim 94\%$) for both acoustic and joint features are much higher than the respective average ($83 \sim 88\%$). This is caused by outliers. Figure 11 shows the box plot of all four metrics of acoustic and joint features. A few outlier classes with very low precision cause low average but do not affect the median. Such outliers are mainly non-human noise classes or human classes with very limited valid samples. When such outliers are excluded, the averages will increase significantly to above $\sim 95\%$.

8.3.2 Performance on New Users

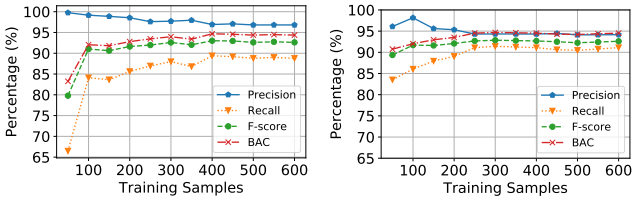
To evaluate how well the pre-trained CNN can extract features for new users, we invite 12 additional volunteers whose data are not used in CNN training. Each volunteer follows the same data collection process for ~ 2 minutes' data, half of which are used for SVM training and the other half for testing. We train a one-class SVM model for each volunteer, and test the model with positive samples from the user and negative samples from all other users, including the data from 50 classes used in CNN training. Table 3 shows that the average precision is over 98%, about 10% increase compared to results in Table 2 due to the absence of outlier classes. Similarly the average recall, F-score and BAC are all improved compared to those in Table 2.

8.3.3 Data Augmentation Evaluation

We evaluate how effective data augmentation can improve the performance by generating "synthesized" training samples when training data is limited. We split 20% samples from the $2min$ data as testing set, and vary the size of training set from 20% to 80%. The data set is shuffled

TABLE 3
Authentication accuracy of new users.

	Mean	Median	Standard Deviation
Precision (%)	98.05	99.21	2.78
Recall (%)	89.36	89.91	1.62
F-score (%)	93.50	94.33	1.68
BAC (%)	93.75	94.52	0.85



(a) Without DA. (b) With DA.

Fig. 12. Classification performance comparison of data augmentation (DA) under different training data amounts.

before the splitting to make it more balanced. Figure 12 shows the precision, recall, F-score and BAC under different amounts of training samples from 50 to 600, which are tested against another 1054 positive testing samples and all the 91708 negative samples from existing 50 classes. It is obvious that data augmentation improves recall significantly, thus F-score and BAC, especially when the training samples are very limited (e.g., <100). As the size grows, the recall with data augmentation is always higher. However the precision decreases to $\sim 95\%$, which is because "synthesized" training samples have more noises, making it easier to have false positives. The performance becomes stable with more than 400 training samples, which can be collected within one minute when registering a new user.

8.3.4 Continuous Modes Evaluation

We evaluate the two continuous modes of presence detection and continuous authentication that uses only acoustics.

Presence Detection. We put the smartphone at different locations as well as holding it in front of the user's face. The detection result is shown on the screen in real-time so that we know the correctness. From our experiments, it can differentiate putting on a table and holding in front of the user with nearly 100% accuracy with unnoticeable delay. Holding in the air sometimes may be detected as user presence when the device is close to some major objects, which may affect timely screen lockup.

Continuous Authentication. To ensure friendly user experience during continuous authentication, a low false negative rate is very important. One volunteer participates in this experiment with a trained model using data when the face is aligned. In the authentication phase, the volunteer keeps using the device as normal and tries to keep it within positions where the face is aligned, with the camera disabled. We evaluate the precision, recall, F-score and BAC when multiple authentication trials are conducted for each cycle. Authentication trial happens every $100ms$ thus one verdict from multiple trials is fast enough, causing no noticeable delay to the user. At least one trial must pass in a cycle to declare authentication success. Figure 13 shows that more trials increase the recall rapidly while decreasing

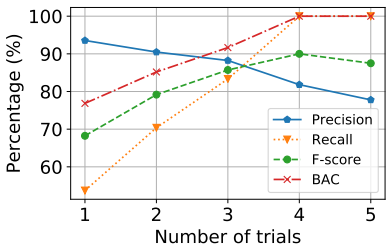


Fig. 13. Continuous authentication performance with different number of trials.

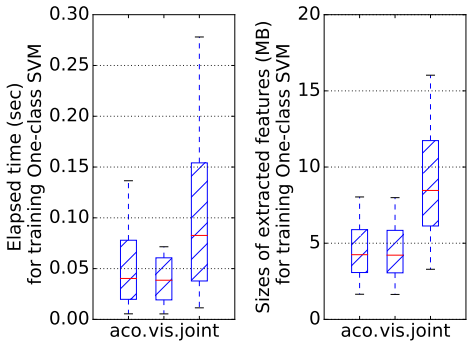


Fig. 14. Elapsed time and size of extracted features for training one-class SVM classifiers using acoustic, vision and joint features.

the precision. This is because more trials give the user more chances to pass, thus reducing denials while increasing false positives. We choose 3 trials for each authentication circle to balance all the metrics.

8.4 Leveraging Sophisticated Visual Features

To evaluate the impact of sophisticated visual features, we invited 10 volunteers whose data are not used for the acoustic training. For each one of them, a $\sim 2mins$ acoustic data and 20 image samples are recorded for evaluation. Since the image data capturing is slower, we populate the image samples by duplicating each sample to meet the number of the acoustic samples according to their timing.

8.4.1 Training Time Cost

To train the authentication model and evaluate the performance, we shuffle and split the collected data into two parts, 80% for training and 20% for testing. From Figure 14, we note that the elapsed time for training SVM with visual features is lower than that with acoustic features, even with the same amount of features. This is because the larger margin between different classes, the shorter time for training SVM with the same regularization parameter C, and the visual feature extractor is trained based on *triplet loss* [27], which encourages clustering the representations of different identities.

8.4.2 Authentication Accuracy

To analyze the impact of acoustic, visual and the combination of both on the overall authentication performance, we compare the above performance metrics using individual features and the joint features. For each user, we use the

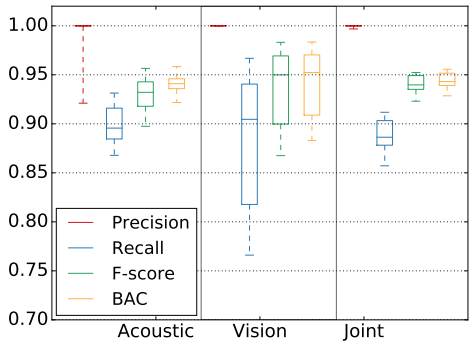


Fig. 15. The precision, recall, f-score and BAC of one-class SVM model using acoustic, vision and joint features.

TABLE 4
Mean/median authentication accuracy of new users with vision, acoustic and joint features.

	Acoustic	Vision	Joint
Precision (%)	98.62 / 100.0	100.0 / 100.0	99.96 / 100.0
Recall (%)	89.83 / 89.56	87.98 / 90.46	88.84 / 88.63
F-score (%)	93.15 / 93.21	93.46 / 94.99	94.07 / 93.97
BAC (%)	94.15 / 94.10	93.99 / 95.23	94.42 / 94.32

positive samples in its test set as positive testing samples and use all the data from other users as negative samples, trying to attack the model. Figure 15 shows the results. As we can see, leveraging acoustic features only, the precision is above 90% with large variances, which is inferior to using visual feature or the joint feature. However, the recall and F-score are significantly better compared to visual features, and slightly better than the joint features. This is because 1) when registering new users, the number of vision samples is limited; 2) when verifying the authentication, joint features examine both modalities in order to produce a positive prediction, which brings down the recall slightly. Thanks to the sophisticated visual features, the precision of joint features are much better than pure acoustic features, demonstrating higher security. Note that, the results are based on the data collected from real human subjects, there are no image/video attacks here. Table 4 shows the authentication results of additional 10 new users with an average precision of 99.96% and F-score of 94.07%. Note that although the precision of joint features is slightly lower than pure visual features, the joint features offer images/videos anti-spoofing capabilities.

8.5 Attack Model Study

Based on our analysis about the attack models in Section 2.1, the replay attack on the facial feature (e.g., image spoofing attacks) is the most concerned adversarial model to our authentication mechanism.

We conduct the image spoofing attack to our system, which is the top concern for existing camera based authentication approaches. We print color photos of 5 volunteers in 10 different sizes on paper, and also display the photos on desktop monitors while zooming in/out gradually, both at various distances between 20 - 50cm to the smartphone. The printed and displayed photos can easily pass the system if

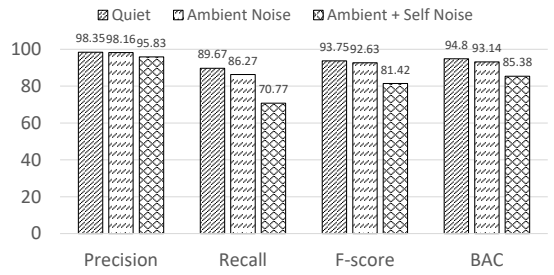


Fig. 16. Performance under difference noises.

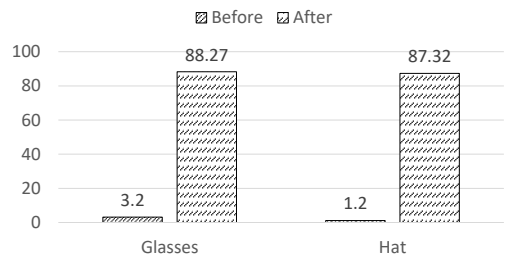


Fig. 17. Average recall of 5 users before/after model updating with new training data.

only vision features are used, but none of them can pass the acoustic or two-factor authentication. This demonstrates the advantage and necessity of the combination of visual and acoustic features.

In addition to the replay attacks on the image feature, we also study the possibility of the replay attacks on the acoustic feature. To capture the acoustic signal for replay attack, we place the eavesdropper around the system when the user performs authentication procedure. We note the presence of a nearby eavesdropper would alter the acoustic response thus producing different features, while a faraway one cannot faithfully capture the acoustic features due to severe attenuation. Therefore, our authentication system is resilient to these attack models.

8.6 Miscellaneous

We evaluate the following factors that have direct impacts on practical use.

User Appearance Changes. Appearance changes such as wearing glasses/hats cause changes in the reflected acoustic signals, thus more false negatives and low recall. To combat such problems, we retrain the SVM model with data samples of new appearances in addition to the existing training data. Figure 17 shows the average recall of 5 users with different appearance changes before/after model update using additional ~ 1 minute's data. Without retraining, the recall drops to single digits. After the retraining, it increases back to normal levels, so correct users can pass easily. This shows retraining is effective combating such changes.

Robustness Against Background Noise. We evaluate the robustness again background noise under different conditions: quiet room, with ambient noise (playing pop music nearby), and with ambient plus self noise (playing music through earpiece speaker on the *same device* during data collection, an extreme condition). Figure 16 shows the results. Except for a slightly lower recall, there is no major difference between quiet and ambient noise conditions, which demonstrates *EchoPrint* is very robust to ambient noise. The ambient plus self noise brings down to recall to $\sim 70\%$, but still the precision remains above 95%.

User Experience Study. We conduct a survey with 20 users (mostly graduate and undergraduate students) to collect their feedback, mainly on two aspects that directly impact user experience, the sensitivity to the emitted sound signal and the effort for new user registration. Out of 20 users, only 4 reported able to hear the high frequency sound from the earpiece when holding the smartphone at a normal distance. Out of 20 users, 9 rated *EchoPrint* equally easy to register as other authentication systems such as image-based face recognition and fingerprint sensor, 6 rated it harder and 5 rated it easier.

8.7 Resource Consumption

We evaluate memory, CPU usage using the Android Studio IDE Profiler tool, and power consumption using Qualcomm's Trepn Profiler tool [34] on Samsung S7 edge, Samsung S8, and Huawei P9.

Memory & CPU Consumption. Table 5 shows the resource consumption on three smartphones when only basic visual features are used. The memory consumption has an average $\sim 22MB$ and maximum $\sim 50MB$, which appears when CNN feature extraction using tensorflow inference is running. The average amount of time for the CPU to complete all the machine learning inferences is low on all phones ($5 \sim 7ms$). The maximum CPU time is around $\sim 30ms$, still very low. Such low memory and CPU usage makes it possible to deploy *EchoPrint* on most existing devices. Table 6 shows the resource consumption when sophisticated visual features are used. In this mode, the image data are streamed to a server for feature extraction, which are returned to mobile device for final authentication. Compared with using basic visual features, it only requires slightly more memory and CPU resources at a cost of possible delay depending on network conditions. This is because of the most heavy computation task – sophisticated visual feature extraction – is offloaded to the server, which makes it possible to maintain real-time recognition on most existing devices.

Response Delay. Response delay is the time needed for the system to produce an authentication result after the raw input signal is ready (Table 5). Samsung S8 has the least delay with an average of $\sim 15ms$, and the other two 32 - $45ms$. The delay approaches maximum when the user keeps moving the phone trying to align the face in the valid area, which incurs a lot of camera preview refreshing and rendering. It is also affected by other computation-heavy background apps. For real-time continuous authentication, the delay between consecutive sound signal emitting is $50ms$. We choose to do authentication every other emitting, leaving sufficient time for processing. We do not evaluate the response delay when sophisticated visual features are used because it highly depends on the network quality. The image data is $\sim 1.4MB$ for each test. Depending on the network quality, the delay time varies. However, unless the wireless networking is highly congested, the delay is acceptable for most use cases.

TABLE 5
Mean/max resource consumption with basic visual features.

Device	Memory (MB)	CPU (ms)	Delay (ms)
$S7$	22.0 / 50.0	6.42 / 31.59	44.87 / 91
$S8$	20.0 / 45.0	5.14 / 29.04	15.33 / 35
$P9$	24.0 / 53.0	7.18 / 23.87	32.68 / 86

TABLE 6
Mean/max resource consumption with sophisticated visual features.

Device	Memory (MB)	CPU (ms)
$S7$	29.3 / 54.9	7.21 / 34.21
$S8$	27.0 / 51.5	6.54 / 31.25
$P9$	32.5 / 61.3	8.37 / 27.63

Power Consumption. We test the three modes and pure vision based authentication using 2D coordinates of facial landmarks, each for 30 minutes to measure power consumption on Samsung S7 Edge, S8 and Huawei P9. We use Qualcomm's Trepn Profiler tool [34], which provides power consumption in mW for a chosen application. We subtract the background power consumption while the screen is on, the increased power consumption caused by different modes are shown in Table 7. The results show that presence detection consumes minimum power, while low power continuous authentication takes less than that of pure light weight vision based authentication. Two-factor authentication has the highest battery consumption; but it is also designed for occasional one-pass authentication finishing in just a few seconds, not long time continuous operation. The slight power increase of vision based mode over LP is due to the simple form of facial landmarks used, which are much lighter weight compared to more sophisticated ones such as those in OpenFace [25].

TABLE 7
Power consumption of different modes.

Device	ULP (mW)	LP (mW)	Two-factor (mW)	Vision (mW)
$S7$	305	1560	2485	1815
$S8$	215	1500	2255	1655
$P9$	265	1510	2375	1725

9 Related Work

Smartphone Authentication. Personal Identification Number (PIN) or a text/graphical password are the earliest and still most widely used smartphone user authentication methods. Despite the simplicity, the PIN or password can be easily peeked by someone close by [4]. Speech recognition is easy to spoof when the voice is recorded, or closely imitated by advanced learning algorithms [35]. BreathPrint [36] senses the user's breath sound, which may change significantly when the user has intense exercises. Vision based face recognition is vulnerable to camouflaged images. Although eye blinks can enhance its security [6], a recorded video can still spoof the system. Fingerprint sensors have achieved great security and convenience. However the sensor takes a lot of precious space, and forging one from fingerprints left by the user is proven practical [8]. More advanced fingerprint sensors use ultrasonics [37] to penetrate the skin and construct 3D imaging, but such sensors are unavailable on most smartphones. Apple's FaceID [10] uses special TrueDepth sensors, bringing extra hardware costs and requiring significant design changes. Unlike all the above solutions, *EchoPrint* is the first to leverage active acoustic sensing combined with visual features for user authentication. It achieves high balanced accuracy ($\sim 95\%$) using existing hardware.

Acoustic-based Face Recognition. Acoustics has been used for face recognition in some prior work [38], [39], [40], [41]. I. E. Dror *et al.* [40] recognize a limited number of five human faces with an accuracy over 96% and the gender of 16 faces with an accuracy of 88% using bat-like sonar input from special ultrasonic sensors. K. Kalgaonkar *et al.* [41] propose a sensing mechanism based on the Doppler effect to capture the patterns of motion of talking faces using ultrasound. K.K. Yoong *et al.* [38], [39] classify up to 10 still faces with an accuracy of 99.73% using hand-crafted features from ultrasound echo signals. Compared to all the above work using special ultrasonic sensors which are not available in consumer electronics, *EchoPrint* uses commodity smartphone speakers and microphones not intended for ultrasonic frequencies. This puts a lot of challenges on the signal design and processing, and much more experiments and tests to find out the best acoustic signal design providing required sensing resolution within hardware limitations, while minimizing the audibility to users. Besides, such prior work uses pure ultrasonic sensing without the aid from vision, thus creating major limitations (e.g., requiring the user to move the head at a fixed location and angle). While *EchoPrint* leverages the vision to align faces using face tracking algorithms for practical two-factor vision-acoustic authentication.

Acoustic Sensing on Smartphones. Acoustic sensing is widely used for distance measurement, thus applications in localization, tracking, stress and encounter detection. BeepBeep [42] measures the distance between two smartphones directly; Liu *et al.* [43] leverage cross-correlation to compute the arrival time difference for keystroke snooping; EchoTag [44] recognizes different locations and BatMapper [16], [45] builds indoor floor plans using echo signals. Besides, acoustic ranging can significantly improve smartphone localization accuracy, e.g., adding constraints among peer phones [46], deploying an anchor network that transmits spatial beacon signals [47], or enabling high-precision infrastructure-free mobile device tracking [48]. FingerIO [49], and LLAP [50] leverage phase shift in received signals for near field finger gesture tracking, achieving $\sim 1cm$ or higher accuracy. ApenaApp [51] monitors the minute chest and abdomen breathing movements using FMCW [52], and SonarBeat [53] monitors breathing beat using signal phase shifts. Compared to them, *EchoPrint* leverages acoustic features from deep neural networks for a different purpose of user authentication.

10 Discussion

Limitations. *EchoPrint* is only a research prototype and far from a well engineered product. It has several main limitations: *i) requirement of face alignment.* To ensure both high true positive and low false negative, face alignment is required

for authentication. It may be inconvenient to users to hold the phone in such positions. In contrast, FaceID has much more flexibility in device holding positions. *ii) limitations from vision. EchoPrint* leverages vision algorithms for facial landmark detection, thus inheriting their limitations. We notice that face tracking is not stable under poor lighting, which makes it hard for face alignment, thus more false negatives. *iii) user appearance changes.* The current CNN feature extractor is trained on limited data, far from exhaustive to be robust against various appearance changes such as hats, glasses or hair styles. Retraining the SVM model with new data is promising for combating such changes. Similar to FaceID, an online model updating mechanism is needed to address such changes dynamically. *iv) continuous authentication usability.* Although we demonstrate *EchoPrint* has the potential for pure acoustic-based continuous authentication, it requires the smartphone aligned and in front of the user's face, which impacts the usability. To mitigate such problem, we can either decrease the frequency of acoustic sampling and authentication at a cost of decreased security, or register users with data from different angles thus enlarging the range of effective authentication.

Future Work. *i) enhancing CNN acoustic feature extractor.* We will collect more training data from more users (e.g., by crowdsourcing) with larger variety. This will further improve performance together with more sophisticated neural network design. *ii) integration with existing solutions.* Due to the rapid progress of deep learning, image based face recognition has achieved unprecedented accuracy. *EchoPrint* can be integrated with existing pure image based solutions to enhance their anti-spoofing capability. *iii) feasibility of replay attacks.* In theory the echoes can be recorded and replayed. However a successful attack is far from trivial: weak, high frequency echoes are difficult to record at sufficient fidelity, and they must be replayed at proper timing after signal emitting. The amount of hardware resources and human efforts needed for successful attacks could be huge, and we plan to find it out. *iv) large scale experiment.* We only have $\sim$ 50 users in the current experiments, while large scale experiment (e.g., thousands or more) is needed for a mature solution. We will seek ways to do experiments at such scale.

11 Conclusion

In this paper, we propose *EchoPrint*, which leverages acoustics and vision on commodity smartphones for two-factor authentication. A convolutional neural network is trained on a large acoustic data set, which is then used as general acoustic feature extractor. Acoustic features are further combined with basic and sophisticated visual features to feed an SVM based classifier for authentication. Experiments show that *EchoPrint* achieves 93.75% balanced accuracy and 93.50% F-score, while the average precision is 98.05% with basic visual features. The precision is further improved to 99.96% with sophisticated visual features.

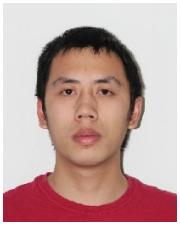
Acknowledgments

This work is supported in part by NSF grants 1652276, 1730291, NSFC 61702035, 62072029, and Beijing NSF L192004.

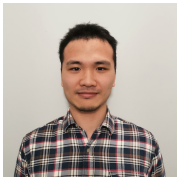
References

[1] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 2018, pp. 321–336.
[2] Google. (2014) Android pay. https://www.android.com/pay/.
[3] Apple. (2018) Apple pay. https://www.apple.com/apple-pay/.
[4] F. Tari, A. Ozok, and S. H. Holden, "A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords," in *Proceedings of the second symposium on Usable privacy and security*. ACM, 2006, pp. 56–66.
[5] N. M. Duc and B. Q. Minh, "Your face is not your password face authentication bypassing lenovo–asus–toshiba," *Black Hat Briefings*, vol. 4, p. 158, 2009.
[6] "Fast face recognition: Eye blink as a reliable behavioral response," *Neuroscience Letters*, vol. 504, no. 1, pp. 49 – 52, 2011.
[7] J. G. Daugman, "Biometric personal identification system based on iris analysis," Mar. 1 1994, uS Patent 5,291,560.
[8] (12/3/2007) How to fool a fingerprint security system as easy as abc. http://www.instructables.com/id/How-To-Fool-a-Fingerprint-Security-System-As-Easy-/.
[9] C. Stein, C. Nickel, and C. Busch, "Fingerphoto recognition with smartphone cameras," in *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, Sept 2012, pp. 1–12.
[10] "About face id advanced technology," https://support.apple.com/en-us/HT208108, 6/8/2018.
[11] "iphone x: This is how much it costs to make one, in components,," http://www.zdnet.com/article/iphone-x-this-is-how-much-it-costs-to-make-in-components/, 11/9/2017.
[12] "Galaxy s8 face recognition already defeated with a simple picture," https://arstechnica.com/gadgets/2017/03/video-shows-galaxy-s8-face-recognition-can-be-defeated-with-a-picture/, 3/31/2017.
[13] J. M. Abreu, T. F. Bastos, and L. Calderón, "Ultrasonic echoes from complex surfaces: an application to object recognition," *Sensors and Actuators A: Physical*, vol. 31, no. 1-3, pp. 182–187, 1992.
[14] (2/28/2014) Background noise reduction: one of your smartphone's greatest tools. https://www.techradar.com/news/phone-and-communications/mobile-phones/background-noise-reduction-one-of-your-smartphone-s-greatest-tools-1229667.
[15] E. C. Ifeachor and B. W. Jervis, *Digital signal processing: a practical approach*. Pearson Education, 2002.
[16] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Batmapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2017, pp. 42–55.
[17] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.
[18] K. G. Derpanis, "Overview of the ransac algorithm," *Image Rochester NY*, vol. 4, no. 1, pp. 2–3, 2010.
[19] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *ISMIR*, vol. 270, 2000, pp. 1–11.
[20] M. Müller, F. Kurth, and M. Clausen, "Audio matching via chroma-based statistical features." in *ISMIR*, vol. 2005, 2005, p. 6th.
[21] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 113–116.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[24] Google. (2017) Introduction to mobile vision. https://developers.google.com/vision/introduction.
[25] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, 2016.
[26] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[27] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
[28] (2020) One-class svm with non-linear kernel (rbf). https://scikit-learn.org/stable/auto_examples/svm/plot_oneclass.html#sphx-glr-auto-examples-svm-plot-oneclass-py.
[29] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.
[30] M. Weber and A. Erdélyi, "On the finite difference analogue of rodrigues' formula," *The American Mathematical Monthly*, vol. 59, no. 3, pp. 163–168, 1952.
[31] F. Chollet *et al.*, "Keras," urlhttps://github.com/keras-team/keras, 2015.
[32] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.
[33] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.
[34] "Qualcomm trepn power profiler," https://developer.qualcomm.com/software/trepn-power-profiler, 2018.
[35] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proceedings of the Sixth International Conference on Learning Representations (ICLR)*, 2018.
[36] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "Breathprint: Breathing acoustics-based user authentication," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 278–291.
[37] Qualcomm, "Fingerprint Sensors," https://www.qualcomm.com/solutions/mobile-computing/features/security/fingerprint-sensors, 2018.
[38] P. McKerrow and K. K. Yoong, "Classifying still faces with ultrasonic sensing," *Robotics and Autonomous Systems*, vol. 55, no. 9, pp. 702–710, 2007.
[39] P. J. McKerrow and K. K. Yoong, "Face classification with ultrasonic sensing," 2006.
[40] I. E. Dror, F. L. Florer, D. Rios, and M. Zagaeski, "Using artificial bat sonar neural networks for complex pattern recognition: Recognizing faces and the speed of a moving target," *Biological Cybernetics*, vol. 74, no. 4, pp. 331–338, 1996.
[41] K. Kalgaonkar and B. Raj, "Recognizing talking faces from acoustic doppler reflections," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.
[42] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, "Beepbeep: A high accuracy acoustic ranging system using cots mobile devices," in *ACM SenSys*, 2007.
[43] 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 *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 2015, pp. 142–154.
[44] K. G. S. Yu-Chih Tung, "Echotag: Accurate infrastructure-free indoor location tagging with smartphones," *MobiCom '15 Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pp. 525–536, 2015.
[45] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Demo: Acoustic sensing based indoor floor plan construction using smartphones," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 2017, pp. 519–521.
[46] H. Liu, Y. Gan, J. Yang, S. Sidhom, Y. Wang, Y. Chen, and F. Ye, "Push the limit of wifi based localization for smartphones," in *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM, 2012, pp. 305–316.
[47] K. Liu, X. Liu, and X. Li, "Guoguo: Enabling fine-grained indoor localization via smartphone," in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. ACM, 2013, pp. 235–248.
[48] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Battracker: High precision infrastructure-free mobile device tracking in indoor environments," in *Proceedings of the 15th ACM Conference on Embedded Networked Sensor Systems (SenSys 2017)*. ACM, 2017.
[49] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "Fingerio: Using active sonar for fine-grained finger tracking," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 1515–1525.
[50] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 2016, pp. 82–94.
[51] R. Nandakumar, S. Gollakota, and N. Watson, "Contactless sleep apnea detection on smartphones," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2015, pp. 45–57.
[52] A. G. Stove, "Linear fmcw radar techniques," in *IEE Proceedings F-Radar and Signal Processing*, vol. 139, no. 5. IET, 1992, pp. 343–350.
[53] X. Wang, R. Huang, and S. Mao, "Sonarbeat: Sonar phase for breathing beat monitoring with smartphones," in *Computer Communication and Networks (ICCCN), 2017 26th International Conference on*. IEEE, 2017, pp. 1–8.



Bing Zhou received his Ph.D. from ECE department, Stony Brook University. He is currently a research staff member in IBM Research. His research interests include mobile computing/sensing, indoor location based services, computer vision and augmented reality. He got a Bachelor from University of Science and Technology of China and Master degree from University of Chinese Academy of Sciences.



Zongxing Xie is a Ph.D. candidate in ECE department, Stony Brook University. His research interests include mobile computing/sensing, smart healthcare. He got a Bachelor from Zhejiang University.

Yinuo Zhang is an MSCS student at Courant Institute of Mathematical Sciences, New York University. Her research interests generally lie in the realm of systems, mainly operating systems and architecture. She received her Bachelor's degree from Stony Brook University.

Jay Lohokare received his Masters of Science in Computer science from Stony Brook University. He is currently a Machine learning Specialist Associate with McKinsey & Company. His research interests include deep reinforcement learning, applied deep learning in HCI and IoT platforms. He got his Bachelors degree from College of Engineering Pune, India.

Ruipeng Gao received the BE degree in communication engineering from the Beijing University of Posts and Telecommunications, in 2010, and the PhD degree in computer science from Peking University, in 2016. He is currently an associated professor with the School of Software Engineering, Beijing Jiaotong University, China. His research interests include wireless communication, mobile computing, and intelligent transportation systems.

Fan Ye received the BE and MS degrees from Tsinghua University, and the PhD degree from the Computer Science Department, UCLA. He is an associate professor in the ECE Department, Stony Brook University. He has published more than 100 peer reviewed papers that have received more than 12,000 citations according to Google Scholar. His research interests include sensing systems, smart aging and health; edge computing; Internet-of-Things; and data centric wireless communication.