This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

Img2Acoustic: A Cross-Modal Gesture Recognition Method Based on Few-Shot Learning

Yongpan Zou, *Member, IEEE*, Jianhao Weng, Wenting Kuang, Yang Jiao, Victor C. M. Leung, *Life Fellow, IEEE*, and Kaishun Wu*, *Fellow, IEEE*

Abstract—Acoustic-based human gesture recognition (HGR) offers diverse applications due to the ubiquity of sensors and touch-free interaction. However, existing machine learning approaches require substantial training data, making the process time-consuming, costly, and labor-intensive. Recent studies have explored cross-modal methods to reduce the need for large training datasets in behavior recognition, but they typically rely on open-source datasets that closely align with the target domain, limiting flexibility and complicating data collection. In this paper, we propose Img2Acoustic, a novel cross-modal acoustic-based HGR approach that leverages models trained on open-source image datasets (*i.e.*, EMNIST, Omniglot) to effectively recognize custom gestures detected via acoustic signals. Our model incorporates a task-aware attention layer (TAAL) and a task-aware local matching layer (TALML), enabling seamless transfer of knowledge from image datasets to acoustic gesture recognition. We implement Img2Acoustic on commercial devices and conduct comprehensive evaluations, demonstrating that our method not only delivers superior accuracy and robustness compared to existing approaches but also eliminates the need for extensive training data collection.

Index Terms—Cross-modal learning, Gesture recognition, Acoustic sensing

1 Introduction

Recently, human gesture recognition (HGR) has gained significant attention and found applications across diverse scenarios. Compared to sensing modalities like radio frequency (RF) signals and inertial measurement units (IMUs), aacoustic sensors have become prominent in HGR due to their robustness and ubiquity, enabling human-computer interaction (HCI) applications such as gesture recognition [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] and texts entry [11], [12], [13], [14], [15], [16]. However, these applications often involve highly costly, time-consuming, and labor-intensive data collection processes. For instance, UltraGesture [4] requires collecting 100 samples per gesture, while Ipanel [12] necessitates 50 samples per gesture. Moreover, models trained on the collected data can only recognize gestures previously known to them.

To reduce the cost of training data collection, recent research has focused on using open-source datasets, including video [17], [18], [19], [20], IMU [21], and WiFi [22] data. These methods either generate target modality data or use data from similar action classes to aid in training. While they achieve good performance, several problems remain: 1) when it comes to video data, researchers need to ensure that actions are not occluded as much as possible; 2) researchers must actively seek specific data that matches the target categories; 3) these systems often lack the flexibility to personalize gestures based on individual needs or preferences. In contrast, we find that image data are more readily available as there are numerous open-source image datasets that can be used. Moreover, there is no need to consider the issue of occlusion when we make use of image datasets. And the data output by acoustic sensors can be represented in the form of images through time-frequency transformation [23]. As a result, inspired by the above points, we put forward such a question: *can we solely leverage the open-source image datasets to train a recognition model for few-shot acoustic gesture recognition?* In response, our proposed Img2Acoustic method emphatically provides an affirmative solution to this question.

Nevertheless, it is not straightforward to realize our goal due to three key challenges. First, it is difficult to identify suitable open-source image datasets to serve as the training set. While there exist many open-source image datasets, not all are suitable for acting as training datasets. This is because different image datasets vary significantly in characteristics, and their relevance to gesture recognition tasks based on acoustic sensing differs widely. A thorough analysis of the target task and careful selection of the appropriate dataset are essential to achieving optimal cross-modal gesture recognition performance. Second, it is challenging to effectively transfer the knowledge learnt from image datasets to our acoustic gesture data, even though suitable datasets have been picked out. This is due to the inherent differences between image data and acoustic signals. Even after converting time-series acoustic signals into spectro-

- *Yongpan Zou, Jianhao Weng, Wenting kuang, and Victor C. M. Leung are with the College of Computer Science and Software Engineering, Shenzhen University, 3688 Nanhai Ave, Shenzhen, Guangdong, China. Email: {wengjianhao2021, kuangwenting2022}@email.szu.edu.cn; yongpan@szu.edu.cn; vleung@ieee.org.*
- *Yang Jiao is with Shenzhen Key Laboratory of Intelligent Bioinformatics, Shenzhen Institute of Advanced Technology, Shenzhen 518055, China. Email: yang.jiao@siat.ac.cn.*
- *Kaishun Wu is with the Information Hub, Hong Kong University of Science and Technology (Guangzhou), 1 Duxue Road, Guangzhou, Guangdong, China. Email: wuks@ust.hk.*

Manuscript received October 16, 2024; revised December 27, 2024.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

grams, the characteristics of the two data types remain significantly distinct. As a result, models trained on image datasets cannot be directly applied to gesture recognition based on acoustic sensing. The last challenge is to enable the trained model to recognize any unseen gestures without retraining. In other words, we aim for the proposed gesture recognition method to identify custom gesture categories based on user preferences and needs, providing an out-of-the-box solution without requiring model retraining.

To tackle these challenges, we first analyze the pattern characteristics of Doppler time-frequency spectrograms produced by gestures. We observe that the key distinguishing feature between different gestures lies in the shape of the frequency shift curves. These patterns closely resemble the characteristics found in handwritten character image datasets. Based on this observation, we select two well-known open-source handwritten character datasets, EMINIST [24] and Omniglot [25], as our training datasets. We propose a novel few-shot learning-based cross-modal acoustic gesture recognition model, named Img2Acoustic. The model integrates two key carefully-designed modules: a Task-Aware Attention Layer (TAAL) and a Task-Aware Local Matching Layer (TALML). The TAAL module adjusts the model's focus on different regions of the feature map. When training with the open-source image datasets EMNIST and Omniglot, TAAL helps the model learn shape information by assigning higher attention weights to regions with distinct shapes, while reducing attention to background areas. The TALML module is based on local descriptor matching. Unlike global descriptors, local descriptors provide richer and more detailed information. Since we train the model on open-source datasets, there is a natural domain gap between the training and testing datasets. TALML effectively reduces this domain gap by allowing the model to concentrate on intra-domain differences. What is more, to prevent the model's feature extractor from excessively focusing on local details, we utilize the output of global features to constrain the output of local features. Finally, we design the overall framework of Img2Acoustic based on the concept of prototype networks, transforming the gesture classification problem into a feature vector matching problem. This approach enables the model to recognize any unseen gestures without re-training, significantly improving its generalization capability.

Based on the above design, we can train Img2Acoustic using only open-source image datasets, enabling it to recognize gestures sensed by acoustic signals. This approach nearly eliminates the need for collecting target modality datasets, significantly improving the efficiency of building gesture recognition systems. Additionally, Img2Acoustic can recognize unseen gestures without retraining, offering strong scalability. We implement the system on Android mobile devices and conduct extensive experiments to evaluate its performance. In a nutshell, the main contributions of this work can be summarized as follows.

- We propose a novel cross-modal acoustic gesture recognition method. This approach leverages only open-source image datasets for model training, significantly reducing the cost of building gesture recognition systems. Additionally, it can recognize unseen, custom gestures based on user preferences and needs, offering strong flexibility.
- We design two novel modules: the Task-Aware Attention Layer (TAAL) and the Task-Aware Local Matching Layer (TALML). These modules combine attention mechanisms, prototype vector generation, and local feature matching techniques, enabling the model to effectively transfer knowledge learned from open-source datasets to the target modality data. To the best of our knowledge, this is also the first work to introduce local feature matching into the field of gesture recognition.
- We conduct comprehensive real-world experiments to evaluate our implementation on Android mobile devices. The results show that Img2Acoustic achieves recognition accuracies of 82.16%, 93.13%, and 95.99% for the 10 handwritten digit gestures from '0' to '9' with 1, 3, and 5 support shots, respectively. Additionally, Img2Acoustic demonstrates excellent performance across various devices and environmental conditions.

The rest of this paper is organized as follows. We discuss the related works in Sec. 2. Sec. 3 and Sec. 4 introduce problem formulation and the proposed method in detail respectively. Sec. 5 gives details of experiments and implementation. Following that, Sec. 6 presents the experimental results. Finally, we discuss the limitations in Sec. 7 and conclude this work in Sec. 8.

2 Related Work

2.1 Cross-modal Solutions in HGR/HAR

To reduce the cost of collecting training data, recent research in HAR and HGR has explored various cross-modal solutions [17], [18], [19], [20], [21], [22], [26]. These studies primarily adopt two technical approaches. On the one hand, some research directly generates target domain data from source domain data for model training. For instance, Vid2Doppler [17] and Midas [20] synthesize millimeter-wave radar data from videos, while IMUTube [18], [19] extracts virtual body tri-axial acceleration from videos for activity recognition, and SignRing [26] extracts finger-level tri-axial acceleration for sign language recognition. However, these approaches require source domain data that precisely matches the activity categories of the target domain, and the activities in the videos must be unobstructed. One the other hand, other studies utilize source domain data to assist in training models for recognizing target domain activities. For example, IMU2Doppler [21] employs readily available smartwatch IMU datasets to aid in training HAR models based on millimeter-wave radar, while RF-CM [22] leverages knowledge from large WiFi datasets to develop HAR systems based on radio frequency signals. Nonetheless, these methods require users to provide a small amount of labeled target domain data for participation in the training process.

In summary, despite the emergence of various cross-modal recognition technologies in HAR, significant shortcomings persist. Methods generating cross-modal data from videos may introduce noise, distortion, or information loss

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

due to factors such as video quality and occlusion, negatively impacting the quality and usability of the datasets. Approaches that leverage source domain data for model training necessitate the collection of a small amount of target modality data for fine-tuning. Furthermore, existing methods require consistency between the categories of the source and target modalities, which can lead to considerable time spent on finding suitable training datasets. Additionally, recognizing unseen categories typically requires model retraining, significantly increasing overhead. In contrast, this paper proposes utilizing a broader range of open-source image datasets for cross-modal gesture recognition in acoustic sensing, thus eliminating the need for category consistency between source and target domains.

2.2 Cross-domain Few-Shot Learning

The goal of few-shot learning is to generalize the knowledge learned from a few labeled samples in auxiliary base classes to new classes with limited labeled samples. Popular research approaches in few-shot learning include prototype-based metric learning [27], [28], [29], meta-learning [30], [31], and transfer learning [32], [33]. Unlike few-shot learning, cross-domain few-shot learning (CD-FSL) aims to transfer knowledge from a source domain to a target domain. The source domain and target domain may possess distinct data distributions. Some recent works [34], [35], [36], [37], [38], [39] make progress on single-source CD-FSL. Among them, the works [34], [35], [36], [37] require fine-tuning the feature extractor to alleviate the domain gap when dealing with a small number of tasks in each target domain. The work [38] involves a complex model that necessitates feature affine transformations on each convolutional layer, making it unsuitable for deployment on mobile devices. The work [39] demonstrates strong domain invariance in local features, but its proposed method limits generalization capabilities, making it unsuitable for use in this work. In comparison, our work draws upon the local feature matching method from the work [40] to implement domain transfer and design a highly real-time and robust system. It is worth noting that while some works, such as the works [32], [33], [34], also use local descriptors, their recognition targets are real images. In contrast, our method focuses on Doppler time-frequency spectrograms which have a uniform background and more consistent feature patterns. As a result, we conduct different operation on prototype vectors to make the model capture subtle differences among support vectors, improving gesture classification accuracy. This design accelerates the matching process and enhances TALML's generalization to diverse gesture spectrograms.

2.3 Acoustic-based HGR/HAR

The widespread use of acoustic sensors in smart devices, due to their contact-free nature, has garnered significant interest among researchers in acoustic-based HCI applications, particularly in gesture recognition [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], text entry [11], [12], [13], [14], [15], [16], and hand tracking [23], [41], [42], [43]. Recent studies [4], [8], [9], [10], [11], [12], [14], [15] have leveraged deep learning networks to enhance the performance of acoustic gesture recognition. However, to improve system generalization, these studies typically require extensive training data. Consequently, encountering new gestures or application scenarios necessitates new data collection, a process that is both time-consuming and labor-intensive. Moreover, training and testing datasets must contain identical categories, preventing the model from recognizing unseen gestures. In contrast, our work introduces two key innovations: first, our method does not require any target modality data for model training, thereby eliminating data collection overhead; second, it enables recognition of unseen gestures without retraining the model.

3 Problem Formulation

In this work, we focus on the N-way K-shot few-shot classification problem [44], where N represents the number of classes and K represents the number of labeled samples for each class. Let $\mathcal{D}_s = \{(\mathbf{X}_i^s, y_i^s)\}_{i=1}^{N_s}$ denote a open-source image dataset, where $\mathbf{X}_i^s \in \mathbb{R}^{H\times W\times D}$ is an image, $y_i^s \in \{1, 2, ..., C\}$ is the corresponding class label, N_s is the number of samples, and C is the number of categories. The model is trained on $\mathcal{D}_s$ with an episodic training mechanism [28]. Let $\mathcal{D}_t = \{(\mathbf{X}_j^t, y_j^t)\}_{j=1}^{N_t}$ denote the acoustic sensing gesture dataset, where $\mathbf{X}_j^t$ and y_j^t are the j-th sample and label, respectively. The model is tested on a series of N-way K-shot episodes randomly sampled in $\mathcal{D}_t$. Note that the classes between $\mathcal{D}_s$ and $\mathcal{D}_t$ are disjoint. In each episode $\mathcal{T}$ (*i.e.*, task), a support set $\mathcal{S} = \{(\mathbf{X}_{S_i}, y_{S_i})\}_{i=1}^{N\cdot K}$ contains a small number of K labeled gesture per class. Besides, a query set $\mathcal{Q} = \{(\mathbf{X}_{Q_i}, y_{Q_i})\}_{i=1}^{M}$ consists of different samples of the same class as $\mathcal{S}$. Here M represents the number of samples in the query set.

The support set is crucial as it provides a limited number of labeled samples for each class, simulating real-world scenarios with sparse data. This small-scale set highlights the model's capacity to learn and generalize from minimal information. Meanwhile, the query set $\mathcal{Q}$ consists of diverse samples from the same classes, serving as an additional validation complementing the support set. Each sample in the query set, denoted as $(\mathbf{X}_{Q_i}, y_{Q_i})$, includes a sample $\mathbf{X}_{Q_i}$ and its corresponding label y_{Q_i}. This design allows the model to learn from the information provided in the support set and test its generalization performance on new samples. In practical scenarios, the user only needs to provide a few support samples per gesture for the system which are collected under predefined conditions, including the environment, device, and angle. Subsequently, the user can interact with the system by making gestures that serve as query samples. The system then matches each query sample with pre-stored support samples to determine the corresponding category. In this stage, the query samples may be collected under various settings with different environments, angles, and ambient noises.

4 System Design

In this section, we present the design details of Img2Acoustic. First, we describe the selection of training dataset and data preprocessing, which involves data augmentation on the open-source image dataset during the

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

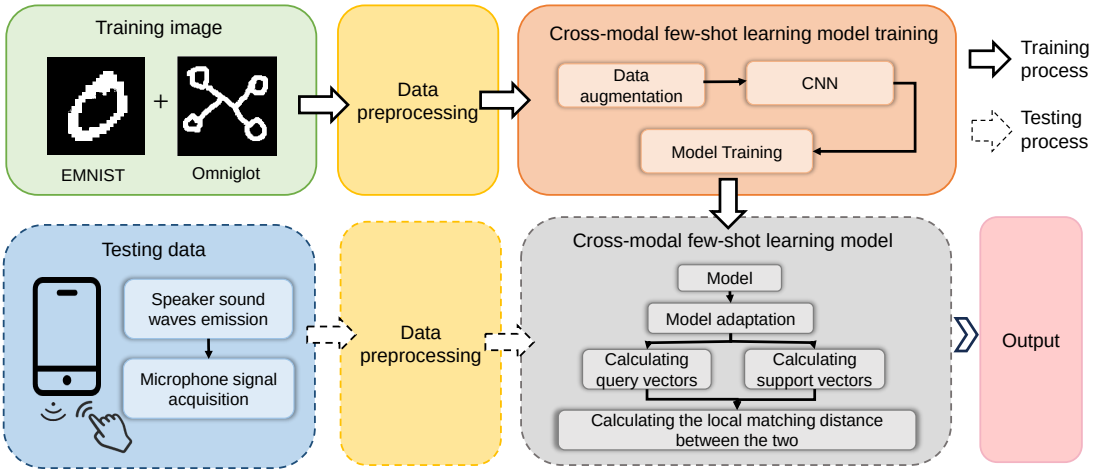


Fig. 1. The system architecture of Img2Acoustic.

(a) Gesture '0' (b) Gesture '1' (c) Gesture '2'

Fig. 2. The spectrograms of different digit gestures.

training phase. Moreover, during the testing phase, time-domain gesture audio signals are transformed into spectrograms. The preprocessed data is used as input for the feature extractor. Then we introduce the proposed model which is composed of a feature extractor and two matching modules. We then detail the classification method employed. Next, we outline the training and testing process of the proposed model. Finally, we summarize the entire pipeline of training and testing the proposed model.

Fig. 1 exhibits the architecture of our system. During the training phase, two open-source image datasets are preprocessed and utilized as input to train the model. Then, the trained model is deployed on an intelligent terminal device with speakers and microphones, and the model does not need to be retrained in this process. When the system is running, the intelligent device will control the speaker to emit high-frequency sound waves that cannot be heard by the human ear. The user can then enter custom gestures through the system on the smart device. After all custom gestures have been entered. In this paper, the obtained gesture data is preprocessed, and then the support set (support vector) of the custom gesture input is obtained through the feature extractor, and is used as the input of the training model. When the user provides all the samples of the custom gesture categories, and then performs gesture operations under the smart device, the executed gestures will go through the same data preprocessing stage. Finally, the feature extractor generates the query sample, that is, the query vector. This vector is locally matched with all the support vectors to get the final gesture category.

4.1 The Selection of Training Datasets

As shown in Fig. 2, through the transformation of STFT, we convert the temporal signals of the gestures collected by the acoustic sensor into time-frequency spectrograms. By examining the unique frequency shift patterns in the spectrograms of different gesture categories, we can effectively distinguish between gesture categories. Hence, we try to utilize the open-source image datasets with obvious shape features to train the model, as shown in Fig. 3. This approach aims to mitigate the domain gap by focusing on prominent shape features. Subsequently, we intend to transfer the trained model to the target domain modality, which can bridge the gap between different modalities and achieve successful recognition.

4.2 Data Preprocessing

4.2.1 *Training data preprocessing*

To make the model more robust and mitigate its tendency to overly emphasize color information from the training dataset, we initially transform binary images into color images. Subsequently, we employ a range of data augmentation methods on the images which involve introducing noise and applying random color variations.

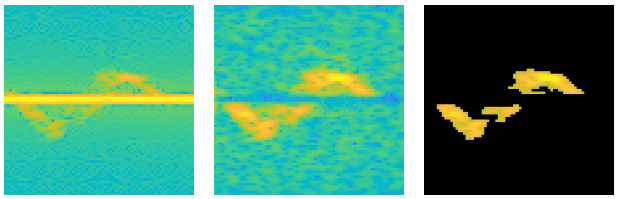
4.2.2 *Acoustic data preprocessing*

The frequency band of environmental noise typically ranges from 1 KHz to 4 KHz [45]. As a result, to avoid interference, we configure the speakers in smart devices to emit a single-frequency sinusoidal ultrasonic wave at 19 KHz. In addition, we also configure the microphone within the same device to capture the echoes reflected from fingers and other objects. First, in order to extract the static component of the center frequency (*i.e.*, 19 KHz), we employ a 3rd-order Butterworth band-stop filter to isolate the bins near the dominant frequency ranging from 18985 Hz to 19015 Hz. Then according to the Doppler effect, the frequency shift caused by finger gestures can be estimated by Eq.(1).

$$\Delta f = f_0 \times |1 - \frac{v_s \pm v_f}{v_s \mp v_f}| \quad (1)$$

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443



(a) Original spectrogram (b) Bandstop filtered spectrogram (c) Enhanced spectrogram

Fig. 3. The spectrograms of the gesture '0' after different preprocessing operations.

where f_0, v_s, and v_f represent the frequency of the emitted ultrasonic wave, the speed of sound in air, and the speed of the finger gesture, respectively. The $\pm$ sign is used when the finger gesture is moving towards or away from the microphone. In this work, the speed of the finger gesture is less than $1\,\mathrm{m/s}$ and the speed of sound in air is about $340\,\mathrm{m/s}$. The resulting frequency shift is approximately $112\,\mathrm{Hz}$. Consequently, the frequency of the received signals is set to $[18800, 19200]\,\mathrm{Hz}$. We apply a 3rd-order band-pass filter to retain the signal within this frequency range. After that, we use the short-time Fourier signals combined with a sliding Hamming window to transform the time-domain signal sequence into a spectrogram. The sliding window has a width of 8192 samples and a stride of 1024 samples. Next, we normalize the time-frequency spectrograms and then enhance them by applying a 2-dimensional Gaussian low-pass filter for smoothing and threshold segmentation. We set the threshold to 0.72. Based on these operations, we can efficiently identify and differentiate the variations within the spectrum of distinct classes. Fig. 3 shows the original spectrogram, the spectrogram after band-stop filtering, and the final enhanced spectrogram, respectively.

4.3 Img2Acoustic Model

To effectively transfer knowledge from the open-source image dataset to our acoustic sensing gesture dataset, we design a model that includes a feature extractor, task-aware attention layer (TAAL), global matching layer, and task-aware local matching layer (TALML). Next, we introduce the design of each module in detail. Fig. 4 shows the overall architecture of the proposed model, where Fig. 4(a) and Fig. 4(b) are the training and testing pipelines, respectively.

4.3.1 Feature extractor

The work [46] shows that shallow neural networks tend to focus on the fine-grained details of the data (*e.g.*, color, shape, *etc.*). Inspired by this, we design a feature extractor f_θ with only four convolutional layers (Conv4) and incorporate a residual structure. This addition enables the feature extractor to more accurately capture the distinct features of the data. Furthermore, we adopt ReLU activation functions for the initial two convolutional layers and GELU activation functions for the subsequent two layers. This decision is motivated by two factors. On one hand, ReLU offers faster computation speed in shallow networks and limits expressive capacity. On the other hand, GELU provides a more intricate information representation capability which is particularly advantageous in deeper networks.

4.3.2 Task-aware attention layer

The task-aware attention layer is used to learn the task-specific attention map, which, in turn highlights the important features of the input image. Based on the inspiration from FiLM [47] and CBAM [48], we design the TAAL module. The TAAL module consists of a Conv Block h_{conv}, followed by a sigmoid function. The Conv Block consists of a convolutional layer with a kernel size of 1, followed by batch normalization and a GELU activation layer. Its purpose is to extract the feature map of the input image. Subsequently, the sigmoid function is applied to normalize the attention map to the range of $[0, 1]$. In the actual testing process, the query samples can only be classified by measuring their distance to the support set. As a result, we modify the model's attention on specific regions of the input image solely based on the support set. Furthermore, before applying the TAAL module, we perform pooling operations on the support set to obtain a comprehensive representation of the support set that changes the feature dimension from $[N, K, D, W, H]$ to $[N, 1, D, W, H]$. First, the support set passes through the TAAL to obtain an attention map. The attention map is calculated by Eq.(2).

$$\mathbf{A} = \sigma(h_{conv}(pool(f_\theta(\mathbf{S})))) \tag{2}$$

where $\mathbf{A}$ is the attention map, σ is the sigmoid function, h_{conv} is the Conv Block and $pool$ is pooling operation.

Then this attention map is multiplied with both the support set and the query set, which adjusts the weights of the feature maps for the support and query sets by Eq.(3). In this way, the model can make specific adjustments to the feature maps of different categories and change the model's focus.

$$\begin{aligned} \mathbf{S}^{'} &= \mathbf{A} \odot \mathbf{S} \\ \mathbf{Q}^{'} &= \mathbf{A} \odot \mathbf{Q} \end{aligned} \tag{3}$$

where $\mathbf{S}^{'}$ and $\mathbf{Q}^{'}$ are the adjusted support set and query set, respectively. $\odot$ is the element-wise multiplication. Finally, the output is a task-specific feature vector obtained from this process action.

4.3.3 Global matching layer

In the global matching layer, we adopt global feature for matching and employ the Euclidean distance as metric function. First, we use the feature extractor f_θ to extract the global feature of the input image. Then we use the TAAL to obtain the task-aware global feature. Next, similar to the prototype network [27], given K-shot episode $\mathcal{T}$, we define the prototype based on Eq.(4). Each prototype is the mean vector of the embedded support points belonging to its class.

$$\boldsymbol{p}_c = \frac{1}{|\mathbf{S}_n|} \sum_{(\mathbf{X}^{'}_{S_i}, y_{S_i}) \in \mathbf{S}^{'}_n} \mathbf{X}^{'}_{S_i}, c \in \{1, 2, ..., N\} \tag{4}$$

where $\mathbf{X}^{'}_{S_i}$ is the attention map obtained by the TAAL module, and $\boldsymbol{p}_c$ is the prototype of the c-th category.

After that, We employ two Conv Block $Conv_g$ to further extract features and reduce the dimensionality of the feature vectors, aiming to learn generic knowledge with image global representation. Finally, we use the Euclidean distance

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

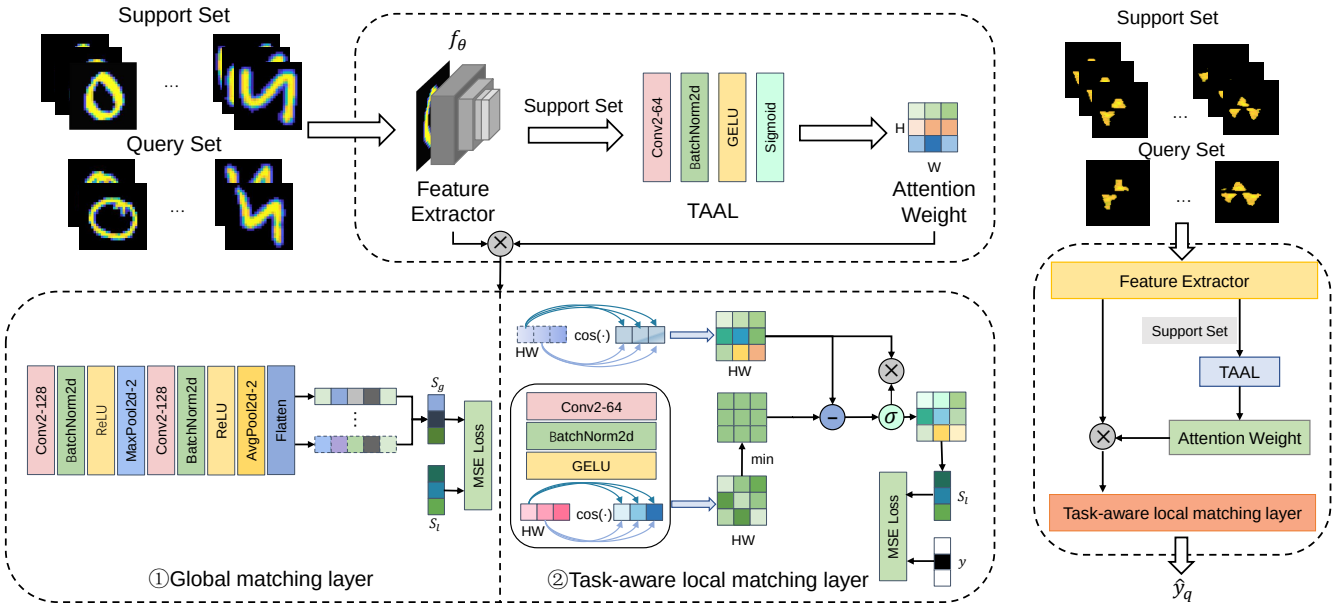


Fig. 4. The network architecture of the proposed model.

to measure the similarity between the query sample and the prototype of each category. The predicted label is the category with the smallest distance. As for the query sample $\mathbf{X}'_{Q_i}$, the global distance metric score S_g corresponding to the support set $\mathbf{S}$ is calculated by Eq.(5).

$$\begin{aligned} \mathbf{G} &= \|\mathbf{X}'_{Q_i} \odot \mathbf{A} - Conv_g(\boldsymbol{p}_c)\|_2 \\ \boldsymbol{S}_g &= softmax(\frac{1}{\mathbf{G}}), \boldsymbol{S}_g \in \mathbb{R}^{1\times N} \end{aligned} \tag{5}$$

where softmax is used to normalize the distance metric score within the range of $[0, 1]$.

4.3.4 Task-aware local matching layer

Recent studies [39], [40] demonstrate that local descriptor-based features surpass global features by offering more granular and detailed information. In the context of acoustic-based gesture recognition, where distinct shape features appear in the spectrogram, leveraging local features for matching can substantially enhance the model's generalization performance, compared to relying on global feature matching. To illustrate this, imagine how human eyes seek discrepancies between two similar images. When we examine the pictures as a whole, it becomes difficult to precisely identify the differences. In contrast, if the comparison is narrowed down to specific local regions, the identification process becomes more straightforward. This analogy aligns with our strategy of transforming the model's recognition approach to favor local classification over global one. In addition, global feature methods frequently perform multiple rounds of feature pooling, resulting in a significant loss of valuable information, particularly when there are significant dissimilarities between the training and testing data. Conversely, the local feature method undergoes fewer pooling iterations, allowing it to preserve more meaningful information. In addition, local feature matching is based on the matching between feature points. Even if there are certain reasons that may cause the displacement of feature points during the process of feature extraction, this is not a concern for local matching. This is because local matching relies on the matching between the current feature point and all other feature points, eliminating the need to worry about this issue. In summary, local features can capture detailed patterns and variations in the data, enable the model to better distinguish different gestures, and perform well on unseen or new examples. This approach enhances the model's cross-domain transferability, which allows it to perform well on gesture spectrogram recognition. By utilizing local features, the model's generalization ability is strengthened, leading to improved effectiveness in recognizing a wide range of gestures. In short, given an image $\mathbf{X}$, after being processed by the previous module, it is expressed as $\mathbf{\Phi}(\mathbf{X}) \in \mathbb{R}^{H\times W\times D}$, which can be regarded as a set of r $(r = HW)$ D-dimension local feature descriptors as

$$\mathbf{\Phi}(\mathbf{X}) = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_r] \in \mathbb{R}^{r\times D} \tag{6}$$

where $\boldsymbol{l}_i$ is the i-th deep local descriptor. In our work, we use an image with a resolution of 84×84 as input, and the output of the previous module is a $19 \times 19 \times 64$ feature map. Therefore, the number of local features is 361, and the dimension of each local feature is 64. First, we normalize the feature vectors of the query set and the support set. The primary objective of normalization is to reduce amplitude discrepancies in feature vectors, ensuring that computations emphasize the directional relationships between them. This serves to prevent inaccuracies in similarity calculations that may arise due to differences in vector magnitudes. Normalization entails dividing each local feature descriptor vector by its respective Euclidean distance, as shown below:

$$\boldsymbol{l}_{norm_i} = \frac{\boldsymbol{l}_i}{\|\boldsymbol{l}_i\|_2} \tag{7}$$

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

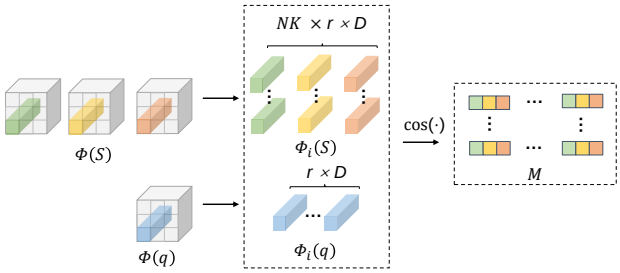


Fig. 5. The diagram of local feature matching.

Following the above, as shown in Fig. 5, we compute the similarity matrix between the feature vectors of the query sample and the support set using cosine similarity. Unlike traditional prototype networks that rely on a single prototype vector for matching, our approach focuses on local matching. In order to enhance the model's domain generalization ability, we perform local matching on all support set images instead of a single prototype vector. Specifically, a given query sample $\mathbf{q}$ will be represented as $\mathbf{\Phi}(\mathbf{q}) = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_r] \in \mathbb{R}^{r \times D}$ and all support images $\mathbf{S}$ will be represented as $\mathbf{\Phi}(\mathbf{S}) = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_{r \times NK}] \in \mathbb{R}^{NK \times r \times D}$. The similarity matrix is calculated by Eq.(8).

$$\begin{aligned} \mathbf{M}_{i,j} &= cos(\Phi_i(q), \Phi_j(S)), i \in [1, r], j \in [1, NK \times r] \\ cos(\Phi_i(q), \Phi_j(S)) &= \frac{\Phi_i(q) \cdot \Phi_j(S)}{\|\Phi_i(q)\| \|\Phi_j(S)\|} \end{aligned} \tag{8}$$

where $\Phi_i(q)$ and $\Phi_j(S)$ represent the i-th point of a query sample and the j-th point of the support set, respectively. $M_{i,j}$ is the cosine similarity between $\Phi_i(q)$ and $\Phi_j(S)$.

We utilize the attention layer to obtain the task-aware local feature for reweighting $\mathbf{M}$. Initially, we process $\mathbf{\Phi}(\mathbf{X})$ through a convolutional block $Conv_l$. This block contains a convolutional layer with a kernel size of 1 and employs the GELU. Next, we construct a new relation matrix $\mathbf{M}'$ using Eq.(8). It's important to note that K is set to 1 in this step. Similar to the prototype network, we generate prototype vectors for each class in the support set using Eq.(4) because they can offer more representative and generalizable representations for classes in the feature space. Furthermore, this approach can help reduce computational time for the model, improving the overall efficiency of the model. Then we use KNN [49] to extract P points from all support samples that are most relevant to each point of the query sample. In this context, P is a hyperparameter and we set it as $N \times H$.

Specifically, a given query sample $\boldsymbol{q}$ will be represented as $Conv_l(\Phi(q)) = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_r] \in \mathbb{R}^{r \times D}$ and all support images S will be represented as $Conv_l(\Phi(S)) = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_{N \times r}] \in \mathbb{R}^{Nr \times D}$. We compute the similarity matrix between the query sample and the support set using cosine similarity by Eq.(8), resulting in $\mathbf{M}' \in \mathbb{R}^{r \times Nr}$. We extract P points from the second dimension of $\mathbf{M}'$ and extract the point with the minimum value v as the threshold to eliminate the noise. Next, through the sigmoid function, we obtain the weights for the reweighted $\mathbf{M}$ using Eq.(9).

$$\begin{aligned} R &= \sigma(\alpha(\mathbf{M}_{i,j} - v)) \\ \mathbf{M}^{new}_{i,j} &= R \odot \mathbf{M}_{i,j}, i \in [1, r], j \in [1, NK \times r] \end{aligned} \tag{9}$$

where α is a hyperparameter, and here we set it as 100. $\mathbf{M}^{new}_{i,j}$ is the reweighted similarity matrix. Due to the cosine similarity calculation, the elements in the matrix $\mathbf{M}_{i,j}$ are within the range of $[-1, 1]$. As a result, it is necessary to introduce a hyperparameter to ensure that the output weights after applying the sigmoid function fall within the range of $[0, 1]$. Furthermore, the sigmoid function is used to ensure that smaller weights tend to approach 0 and larger weights tend to approach 1. In this way, we can provide more attention to values with high similarity and less attention to values with low similarity without completely assigning a value of 0 to those with low similarity. Finally, we calculate the local similarity scores S_l between the query sample and a sample of each support set by Eq.(10), and then determine the final category of the query sample based on these scores.

$$\begin{aligned} S_l &= \frac{1}{r} \sum_{i=1}^{r} \sum_{j=1}^{K \times r} \mathbf{M}^{new}_{i,j} \\ \hat{y_q} &= \arg max(S_l), \hat{y_q} \in [1, N] \end{aligned} \tag{10}$$

where y_q is the predicted category of the query sample.

Algorithm 1 Training procedure of Img2Acoustic

Require: Training dataset $\mathcal{D}_s$, number of episodes T, number of way N, number of support images per class K_S, number of query images per class K_Q, learning rate η, number of training epochs E.
Ensure: Trained model M.

```
randomly initialize the model M parameter θ.
for e = 1 to E do
  for n = 1 to T do
    Sample N classes from D_s.
    Sample K_S support images and K_Q query images from each class.
    Extract features for all images using the feature extractor.
    Using support images Calculate the task-specific attention map A using Eq.(2).
    Using the task-specific attention map A to adjust the weights of the feature maps for the support images and query images by Eq.(3).
    Calculate the global similarity matrix G using Eq.(5).
    Calculate the local similarity matrix M_l using Eq.(9).
    Calculate the global similarity score S_g using Eq.(5).
    Calculate the local similarity score S_l using Eq.(10).
    Calculate the total loss L_T using Eq.(11).
    Update the parameters of the model M using L_T.
  end for
end for
return optimized Parameter θ_*.
```

4.4 Feature Measurement

To address the issue of categories inconsistency between the training dataset and the testing dataset, we adopted a metric learning strategy similar to that presented in [27]. This method diverges from traditional models that use fully connected layers for classification. Instead, it employs a

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

metric learning approach to measure the similarity between feature vectors of different samples. Detailed metrics classification process is discussed in Sec. 4.3.4.

4.5 Training and Testing

In the training phase, we use episodic training along with mean squared error (MSE) loss to train Img2Acoustic. We can obtain the similarity scores S_g and S_l from the global branch and the local branch, respectively. In summary, for a few-shot episode $\mathcal{T}$, the total loss function of the training phase is as follows:

$$\begin{aligned}\mathcal{L}_T &= \mathcal{L}_{g-l} + \mathcal{L}_l \\ &= \sum_{i=1}^{I}(S_{g_i} - S_{l_i})^2 + \sum_{i=1}^{I}(S_{l_i} - y_i)^2\end{aligned} \tag{11}$$

where I is the number of query images in the episode $\mathcal{T}$, y_i is the ground truth of the i-th query image.

On the one hand, We constrain the difference between the similarity scores of the global branch and the local branch through $\mathcal{L}_{g-l}$. This approach is primarily aimed at preventing the model from overly fixating on local information at specific locations. By implementing this approach, the model becomes capable of matching local features within a context of globally abundant and informative features. This strategy ensures a balanced exploration of local and global cues, enhancing the model's overall performance in feature matching tasks. On the other hand, we use $\mathcal{L}_l$ to constrain the difference between the prediction results of the local branch and the ground truth. The training procedure of Img2Acoustic is shown in Algorithm(1).

In the testing phase, we first utilize the feature extractor for each image. And then we calculate the task-specific attention map A using Eq.(2). Next, we use A to adjust the weights of the feature maps for the support images and query image. Finally, when both global and local branches contribute to category predictions jointly, we are unable to ascertain the weighting between the global and local scores. Given that the feature information extracted by the global branch may not be sufficiently detailed, we rely solely on the local branch to generate the prediction result for the queried image. It is worth noting that this method does not require fine-tuning the model during the testing phase on the target domain.

5 Experiments

5.1 Datasets

5.1.1 Training datasets

In this work, we can effectively distinguish gesture categories by relying on the shape of the frequency shift in the time-frequency spectrogram transformed by the Short-time Fourier Transform of the acoustic sensing gesture datasets. Therefore, we consider training the model using datasets with significant shape variations between categories. We use the following image datasets to train the model:

- **EMNIST** [24]: EMNIST is a dataset of handwritten digits and letters. It consists of 731668 training images. Each one is a 28×28 binary image. In order to avoid some of the categories being more similar to each other, we picked 10 basic digits from '0' to '9' and 26 English letters from 'A' to 'Z' as the categories for training the model.
- **Omniglot** [25]: Omniglot is a dataset of handwritten characters. It consists of $1,623$ characters from 50 different alphabets. Each image is a 105×105 binary image.

All the images are resized to 84×84 pixels and converted to color ones.

5.1.2 Testing datasets

In this work, we primarily utilize a dataset of handwritten digit gestures (0–9) collected via an Android application we developed to evaluate Img2Acoustic's performance. The application controls the device's speaker to emit 19 KHz sinusoidal audio signals and simultaneously uses the microphone to capture echoes at a 44.1 KHz sampling rate. We select digit gestures for performance evaluation due to several key reasons. First, their predefined and standardized trajectories provide a more objective and reliable basis for system assessment compared to self-defined gestures, which might favor easier recognition. Second, the fixed set of digit gestures ensures consistent and fair comparisons across systems. Finally, digit gesture recognition has broad practical applications, such as dialing phone numbers, entering PIN codes, and beyond. By assigning digit gestures to specific commands, this method also extends to a wide range of human-computer interaction scenarios.

We recruit 10 participants from our university, consisting of 6 males and 4 females, to participate in the experiments. Each participant is asked to perform a hand-written digit gesture 10 times. To ensure a comprehensive evaluation, we consider several key conditions and categorized all potential experimental setups based on different combinations of these factors. This allows us to thoroughly assess the impact of various variables on the experiment's outcomes. Table 1 outlines the experimental setups used in this study. The default scenario was used for model performance evaluation, while other scenarios were applied for robustness assessment. Fig. 6 illustrates the writing conventions for the 10 digit gestures. In addition, we also collect 26 letter gestures and 8 hand gestures for evaluation. The data collection process for these two datasets follows a similar approach to the default scenario. A detailed description of these datasets is in Sec. 6.8.

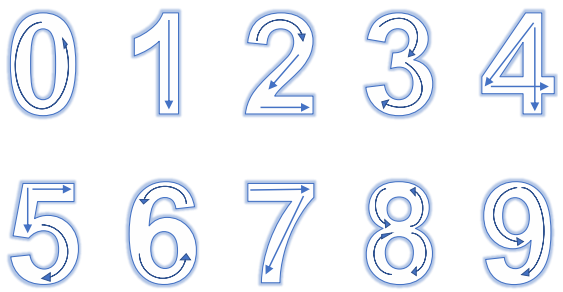


Fig. 6. Writing conventions for the 10 digit gestures.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

TABLE 1
Experimental setups

Evaluation		Data Samples Collection
Overall Performance		Default Scenario: 45 dB to 55 dB, Angle 0°,0 cm, Samsung Galaxy Tab S2 10 Users × 10 Gestures × 10 Repetitions = 1000
System Robustness	2 Environments	1. 45 dB to 55 dB: 10 Users × 10 Gestures × 10 Repetitions = 1000
		2. 55 dB to 65 dB: 10 Users × 10 Gestures × 10 Repetitions = 1000
	2 devices	1. Samsung Galaxy Tab S2: 10 Users × 10 Gestures × 10 Repetitions = 1000
		2. Xiao Mi Mix2: 10 Users × 10 Gestures × 10 Repetitions = 1000
	4 speaker sound levels	70 dB/60 dB/50 dB/40 dB: 4 speaker sound levels × 5 User × 10 Gestures × 10 Repetitions = 2000
	5 angles	-15° /-7.5° /0° /7.5° /15°: 5 angles × 5 User × 10 Gestures × 10 Repetitions = 2500
	5 distances	0 cm/5 cm/10 cm/15 cm/20 cm: 5 distances × 5 User × 10 Gestures × 10 Repetitions = 2500

5.2 Evaluation Metrics

As our primary concern is the system's ability to accurately recognize gestures performed by a user, we chose accuracy metric as the criterion for evaluation. Accuracy is calculated using Eq.(12).

$$Accuracy = \frac{Count_{Pred=True}}{Count_{All}} \times 100\% \tag{12}$$

where $Count_{Pred=True}$ is the number of correct predictions, and $Count_{All}$ is the total number of predictions.

Furthermore, to assess the concentration level of the testing results, we calculate the standard deviation (STD) of the data by Eq.(13).

$$STD = \sqrt{\frac{\sum_{i=1}^{n}(\mathbf{x}_i - \overline{x})^2}{n-1}} \tag{13}$$

where x_i is the accuracy of the i-th task, $\overline{x}$ is the mean of the all task, and n is the total number of all task.

5.3 Baselines

We compare the proposed Img2Acoustic model with four baseline methods closely related to our work. ProtoNet [27] employs a training approach that aligns with our model's few-shot learning strategy, utilizing cutting-edge techniques in the field. DN4 [40] shares similarities with our local matching concept, which we build upon to enhance the network's performance for our application. LDP-net [39], a cross-dataset few-shot learning method and a refinement of ProtoNet, parallels our approach by integrating both local and global feature representations.

- **No Train**: We directly use the model architecture shown in Fig. 4 without being pre-trained on open-source image datasets for testing. The goal is to determine whether training on these datasets enhances the model's effectiveness.
- **ProtoNet**: Prototypical Network (ProtoNet) [27] is a cutting-edge few-shot learning algorithm based on the principles of meta-learning. ProtoNet constructs prototypes in the embedding space when a small amount of training data is provided. Each prototype serves as a representative for a specific class. In the inference phase, ProtoNet employs the Euclidean distance metric to classify new samples by associating them with the nearest prototype, thus determining their respective categories.
- **DN4**: Deep Nearest Neighbor Neural Network (DN4) [40] is a few-shot learning algorithm that uses a deep neural network to learn the optimal deep local descriptors for the image-to-class measure.
- **LDP-net**: Local-global Distillation Prototypical Network (LDP-net) [39] revisits the reasons behind the limited cross-domain performance of prototype networks. It establishes a two-branch network, utilizing local features to distill global features.

In the implementation of these baselines, ProtoNet and DN4 make use of the same feature extractor as our proposed model in order for fair comparison. Since the LDP-net model requires pre-training, training without pre-training hampers the model's ability to effectively optimize the loss, leading to degraded performance. Therefore, LDP-net specifically uses ResNet10 as its feature extractor. In this paper, we directly use the pre-trained model parameters provided by this work as the training parameters for the LDP-net model and fine-tune them using our training datasets.

5.4 Implementation Details

We use the Adam optimizer with a learning rate of $1e-3$. The model is trained for 30 epochs. In each epoch, we randomly sample 100 episodes from the source domain. In each episode, we set the number of classes to 5, the number of support samples of each class to 5, and the query sample size of each class to 10. Since lacking a validation set for model selection, we utilize the checkpoint saved after the last epoch as the final model. The data processing coded in Python runs on a computer (Intel i9-10900K CPU @ 3.70 GHz, 128 GB memory). The model training is conducted with PyTorch library (Python 3.8) on a server with NVIDIA RTX A6000 GPU.

5.5 Evaluation Protocol

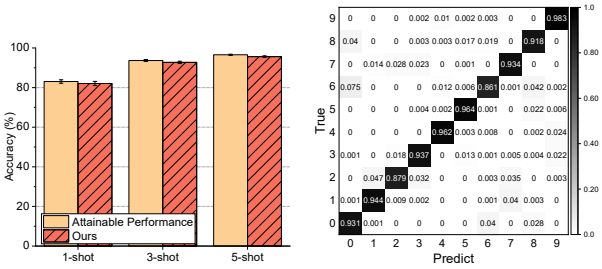
We evaluate the proposed method using acoustic gesture datasets. For each target domain, we randomly sample 500 N-way K-shot, with each class containing 5 query samples. We compute the average accuracy across these sampled tasks. In all validation experiments, K is set to 1, 3, and 5. Unless otherwise noted, the environment for the provided support samples is maintained under the default conditions.

6 PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed Img2Acoustic model.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443



(a) The recognition accuracies of 10 digit gestures with different numbers of shots.

(b) The average confusion matrix of 10 digit gestures over different participants with 3 shots.

Fig. 7. The overall performance of Img2Acoustic.

6.1 Overall Performance

We evaluate the overall performance of the proposed Img2Acoustic model on the default scenario. Specifically, we train the network using both the EMNIST and Omniglot datasets and then conduct testing under the default scenario. When each participant provides 1, 3, and 5 shots for each gesture category, as shown in Fig. 7(a), Img2Acoustic achieves averaged accuracies of 82.16%, 93.13%, and 95.99% in the default scenario. The confusion matrix of 10 classes of digit gestures averaged across 10 participants is shown in Fig. 7(b). From the confusion matrix, we can observe that digits '6' and '1' are likely to be misclassified as '0' and '7' with less than 1% probability. This primarily happens due to the similarity in gestures between these two digits, which leads to confusion in the spectrograms obtained from the sensor.

Futhermore, we train the model on an acoustic sensing dataset containing 26 classes of letter gestures, and subsequently test the model using another acoustic sensing dataset that includes 10 classes of digit gestures. The purpose is to ensure that the training and testing sets share the same modality, thereby testing the model's potential upper limit and providing a more comprehensive reflection of our model's performance. We consider the accuracy obtained by this approach as our attainable performance. When provided with 1, 3, and 5 shots per class, we achieve recognition accuracies of 83.06%, 93.68%, and 96.54% respectively for recognizing 10 classes of digit gestures. The results demonstrate that the proposed Img2Acoustic model achieves accuracy comparable to the attainable performance. This highlights the effectiveness of the model in recognizing gestures in the target domain, even without any specific training data in that modality. Furthermore, the model demonstrates its ability to effectively transfer knowledge from open-source image datasets to acoustic gesture datasets.

6.2 Ablation Study

To gain a deeper understanding of the network design, we conduct ablation experiments to assess the impact of each component. These mainly include the loss function that constrains the relationship between the local branch and the global branch (LG) and the TAAL module. The results of the ablation study are shown in Fig. 8. The findings demonstrate that the enhanced Img2Acoustic model, which includes the TAAL module and a loss function that constrains the relationship between the local and global branches, outperforms the Img2Acoustic model without these features. Particularly, in the 1-shot scenario the average accuracy of our model increases by 4.23%. The TAAL module plays a role in effectively learning the task-specific attention map, which contributes to the model's generalization ability. And the utilization of a loss function that constrains the relationship between the local branch and the global branch further enhances the model's performance in terms of generalization. Furthermore, we conduct a validation to confirm the efficacy of the TALML module. In order to evaluate the impact of the TALML module on the prediction, we conduct an experiment excluding this module. We also train the model on the output of the global features and align them with the actual labels to calculate the losses, a process similar to ProtoNet in Sec. 5.3, but with an additional TAAL module. The result indicate a notable decrease in model performance, averaging a 17.08% decrease across various shot scenarios when the TALML module is removed. We anticipate that the main reason is that the model, during the training process of the local matching module, learns the matching relationship between image feature points rather than the relationship between the entire image feature, which can reduce the domain difference between the open-source image datasets and the acoustic sensing gesture dataset. In this scenario, both training and testing exclusively utilize the global matching layer branch from Fig. 4(a). This result further proves of the effectiveness of the TALML module we design. Finally, we verified the impact of data augmentation on the model performance. The results indicate that when the training data is not augmented, the model's performance decreased by an average of 1.67% in scenarios with 1, 3, and 5 support samples. This further validates the effectiveness of the data augmentation method proposed in this paper for model training.

6.3 Baseline Comparison

In this part, we assess the performance of Img2Acoustic against other baseline methods described in Sec. 5.3 with our acoustic-based gesture dataset. Fig. 9 shows the average recognition accuracies of the proposed Img2Acoustic model and the baseline methods. We can see that Img2Acoustic outperforms all the baseline methods. Particularly in the 1-shot scenario, Img2Acoustic demonstrates an average accuracy improvement of 10.18% compared to the best baseline method. We speculate that the improvement in performance is mainly attributed to the design of the TAAL, TALML modules, and the loss function LG. These components enable our model to better weight the changing characteristics of different input tasks, thereby adjusting the model's focus on specific feature regions and consequently enhancing overall performance. Besides, LDP-net is designed for cross-dataset applications, and its performance on our dataset does not meet our expectations. One of the possible explanations is the similarity among gesture categories in our datasets, and it poses challenges when relying solely on global features for prediction. We are also astonished to find that our model outperforms some trained baseline methods even without formal training. We surmise that this could

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

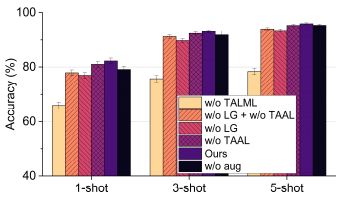


Fig. 8. The results of ablation study.

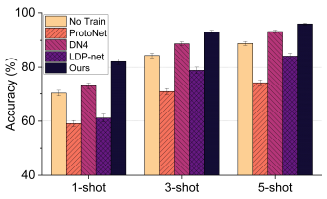


Fig. 9. The comparison of different methods.

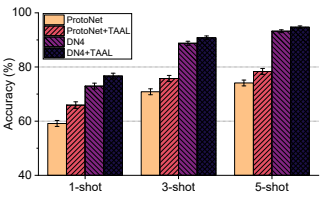


Fig. 10. The impact of TAAL.

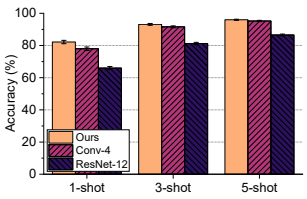


Fig. 11. The comparison of different feature extractors.

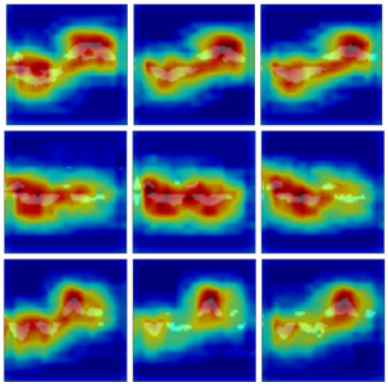


Fig. 12. The visualization of feature maps.

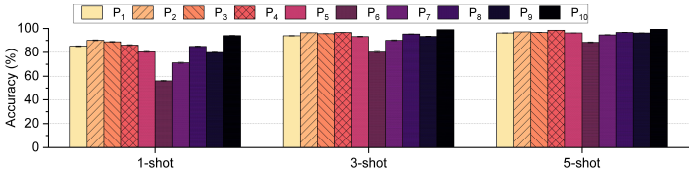


Fig. 13. The recognition accuracies of digit gestures for different participants $P_1 \sim P_{10}$.

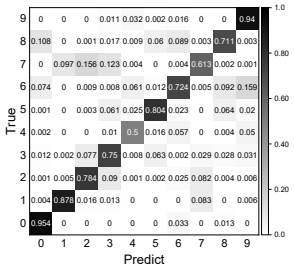


Fig. 14. P_6's confusion matrix for digit gestures recognition.

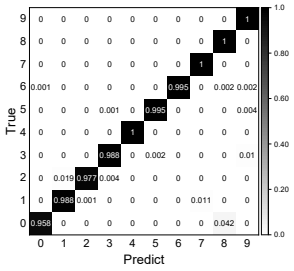


Fig. 15. P_{10}'s confusion matrix for digit gestures recognition.

be attributed to the distinctive nature of our dataset, where the black background highlights the segments intended for recognition. In conclusion, by leveraging our local matching module, which performs correlation matching point by point, we are able to achieve a relatively high accuracy even without formal training.

6.4 The Performance of TAAL

Due to the flexibility of Task-Aware Attention Layer (TAAL) in being applied to partially relevant baselines, we further assess the effectiveness of integrating TAAL into different baselines. Specifically, TAAL is added after the feature extraction module. Similar to Sec. 4.3.2, the testing process solely utilized the support set to generate attention weight maps. The testing results are depicted in Fig. 10. From the results, it is evident that the effectiveness of the designed TAAL module. Across varying numbers of support samples, the performance with the inclusion of the TAAL module consistently outperformed that without it.

6.5 The Selection of Feature Extractor

We further assess the performance of the improved feature extractor and conducted a detailed comparison with the classic ResNet-12 feature extractor and the unimproved Conv-4. The parameter counts for the three types of feature extractors were 0.18 M, 12.42 M, and 0.11 M, respectively. In the comparative experiments, only the feature extractor module is altered, while all other settings remained unchanged. The final testing results are shown in Fig. 11.

From the results, it is evident that our designed feature extractor module outperforms the others. Specifically, in the 1-shot scenario, the performance of the improved feature extractor module surpassed that of the unimproved Conv-4 by 4.17%, demonstrating the effectiveness of our enhancements. However, the recognition performance of ResNet-12 was relatively poor, consistent with the analysis in Sec. 4.3.1 of this paper. The use of a more complex feature extraction module such as ResNet-12 may lead the network to extract deeper semantic information rather than the shallow semantic information intended in our study.

6.6 Visualization of Feature Map

The Class Activation Map (CAM) [50] is a visualization technique utilized to elucidate the key areas of focus in the image classification process undertaken by a deep learning model, thereby shedding light on the decision-making process. This paper presents the CAM of acoustic gesture data processed by the feature extractor as shown in Fig. 12. It can be found that the model performs well in distinguishing the region of interest. However, in the process of feature extraction, the refinement of the model is obviously insufficient. In order to classify the features output by the feature extractor more accurately, it is a useful solution to introduce a local matching method. Through local matching, the model can focus more accurately on the region of interest, improve the refinement of feature matching, and enhance the recognition performance.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

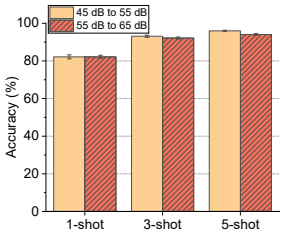


Fig. 16. The recognition performance of digit gestures in different environments.

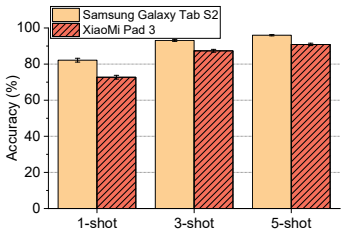


Fig. 17. The recognition performance of digit gestures with different devices.

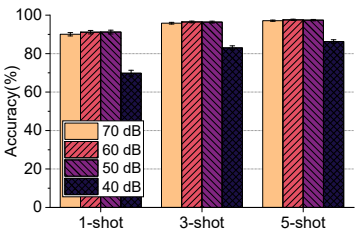
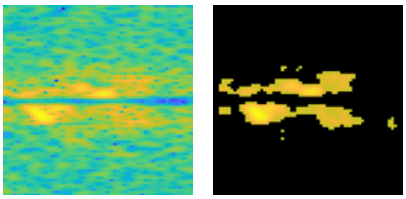
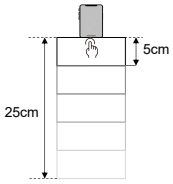


Fig. 18. The recognition performance at different sound levels.

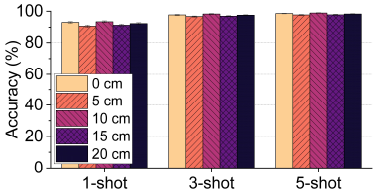


(a) Before data preprocessing. (b) After data preprocessing.

Fig. 19. The spectrograms of a gesture before and after preprocessing at 40 dB.



(a) Initial writing position at different distances.



(b) The recognition accuracies of Img2Acoustic at different distances.

Fig. 20. The impact of distance on recognition performance.

6.7 The robustness of Img2Acoustic in the field

6.7.1 Impact of people

Fig. 13 shows the recognition accuracies of the proposed Img2Acoustic model under different participants. When providing 3 shots per class, participants exhibit varying recognition accuracies evidently, ranging from 80.17% (P_6) to 99% (P_{10}). To illustrate the misclassification of participant P_6, we generate the confusion matrix for the aforementioned test results, as depicted in Fig. 14. In participant P_6's digit recognition environment, the digit gestures '6', '8', '0', and '4' exhibit a tendency to be confused with each other. This confusion primarily arises from the similarity in frequency-shifting features generated by the handwriting motions of these four digits. Failure to distinctly mark the beginning and ending motions while writing may result in similarities between these digits. For example, the digit gesture '7' differs from '1' with the presence of a horizontal stroke. In addition, participant P_6's shorter horizontal stroke in writing '7' which contributes to the confusion between these two digits. Likewise, the beginning and ending movements of the number '2' are unclear and easily confuse with '7'. The accuracy rates of some participants are lower in the digit recognition task with 3 shots. However, when participants provide 5 shots, the average accuracy rates significantly increase to around 90% or higher. Furthermore, we showcase the top-performing participant, P_{10}, in Fig. 15. In summary, the misclassified categories primarily result from the gestures' tendency to be confused by similar actions.

6.7.2 Impact of environment

To evaluate the robustness of the proposed Img2Acoustic model in the real scenarios, we conduct a field study with 10 participants. Each participant is asked to perform 10 classes of digit gestures in 2 different environments, including a quiet room (45 dB to 50 dB) and a noisy room (55 dB to 65 dB). The recognition accuracies of the proposed Img2Acoustic model in these two environments are shown in Fig. 16. The results demonstrate that the proposed Img2Acoustic model can effectively recognize gestures in different environments. This is mainly attributed to the utilization of 19 KHz high-frequency acoustic signals as a perceptual medium in this study, which significantly exceeds the typical levels of ambient noise in regular environments. In daily life, noises are generally present at frequencies below 1000 Hz.

6.7.3 Impact of speaker sound levels

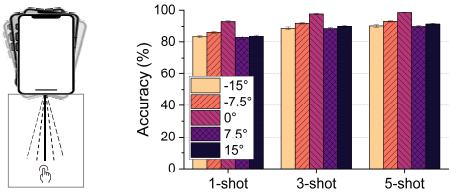
Due to the varying volume levels of users' daily smartphones, we evaluate the proposed model under different sound pressure conditions using the same speaker. In the previous device (Samsung Galaxy Tab S2) that we use, the maximum speaker sound pressure is around 60 dB, while mainstream smartphones on the market typically have sound pressures of 70 dB or higher. Therefore, we select a new experimental device, *i.e.*, Samsung Galaxy S20, and conduct experiments at four different sound pressure levels.

We measure sound levels by placing a sound level meter under the smartphone speaker. After measurement, we determine the corresponding volume values for the four sound pressure levels as follows: 70 dB corresponds to the maximum volume, 60 dB corresponds to 66.67% of the maximum volume, 50 dB corresponds to 40% of the maximum volume, and 40 dB corresponds to 20% of the maximum volume. The recognition accuracies under these four sound pressure levels are shown in Fig. 18.

It can be observed that the model's performance impact is negligible in the 70 db, 60 dB and 50 dB scenarios. However, in the 40 dB scenario, the model's accuracy significantly decreases. This may be due to the decrease in speaker emission power as the sound pressure decreases, leading to a reduction in the energy of the echoes collected

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443



(a) Device at different angles. (b) The recognition accuracies of Img2Acoustic at different angles.

Fig. 21. The impact of device angle.

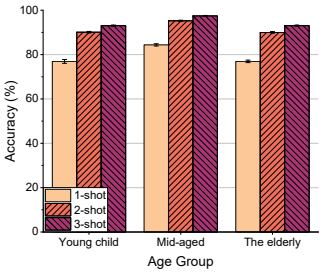


Fig. 22. The recognition accuracies of different gesture datasets.

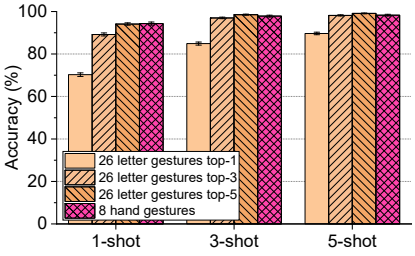


Fig. 23. The recognition accuracies of different gesture datasets.

by the microphone. The frequency variations induced by gestures can easily blend with the inherent noise of the device, leading to a reduction in the signal-to-noise ratio of the signal. Consequently, during data preprocessing, it becomes challenging to distinguish effective signals from the noise, as depicted in Fig. 19. Since the maximum volume of speakers on current smart devices is usually more than 60 dB, our system can achieve good performance on most commercial devices.

6.7.4 Impact of device

To evaluate the robustness of the proposed Img2Acoustic model on different devices, we conduct a field study involving 10 participants. Each participant is asked to perform 10 classes of digital gestures on 2 different devices, including Samsung Galaxy Tab S2 and Xiaomi Mix2. The recognition accuracies of the proposed Img2Acoustic model on these two devices are shown in Fig. 17. Compared to the Samsung Galaxy Tab S2, the Xiaomi Mix2 exhibits slightly lower average recognition accuracy. This is mainly because the background noise introduces interference in the high-frequency components of the spectrograms, which affects the accuracy of recognition. However, when these participants provide 5 shots, the average recognition accuracy of the proposed Img2Acoustic model on both devices is above 90%, which demonstrates the robustness of the proposed Img2Acoustic model on different devices.

6.7.5 Impact of distance

Due to variations of users' writing habits and the decay of acoustic signals over distance during transmission, we enlisted five participants to evaluate the impact of device distance on our system. These participants were instructed to perform gestures at five different distances to accommodate the diverse daily writing habits of users. It's important to clarify that 'distance' here specifically refers to the spatial interval between the initial writing position of the experimenter and the device, as depicted in Fig. 20(a).

The five distance ranges tested in the experiments include $[0, 5]$ cm, $[5, 10]$ cm, $[10, 15]$ cm, $[15, 20]$ cm, and $[20, 25]$ cm. The recognition accuracy of the proposed model at different distances is illustrated in Fig. 20(b). It is observed that different writing start positions have little effect on model performance. This validates that the system designed in this paper can effectively accommodate the daily writing gesture habits of diverse users.

(a) Gesture '0' at 0°. (b) Gesture '0' at 7.5°. (c) Gesture '6' at 0°.

Fig. 24. Confusing spectrograms in different situations.

6.7.6 Impact of angle

Considering that users may experience slight changes of device angle during gesture interactions, we aim to assess the impact of device angles on system performance. Working with the same participants as in Sec. 6.7.5, we instruct the participants to perform gestures at five different angles. The variations in device angles are shown in Fig. 21(a). The five different angles are $-15°$, $-7.5°$, $0°$, $7.5°$, and $15°$. The recognition performance of the proposed Img2Acoustic model at different angles is shown in Fig. 21(b). We can see that in the case of 1-shot, the variation of the device angles has a noticeable impact on system performance. This occurs primarily due to the similarity in frequency deviations among certain gesture categories from specific users. Even minor changes in device angle lead to alterations in the original relative motion trends, thereby amplifying the similarity between different categories. As depicted in Fig. 24, When the device is rotated $7.5°$ to the right, the frequency deviation of some users' gesture '0' will change to a certain extent. This is mainly because the user's finishing movement when writing '0' changes from being closer to the device to being far away from the device, resulting in Fig. 24(b), there is an downward frequency shift at the end. Moreover, when users provide 5-shot, the average recognition accuracy across the five angles exceeds 90%. This underscores the robustness of Img2Acoustic across various angles.

6.7.7 Impact of users' ages

Users from different age groups exhibit distinct gesture patterns, which may influence recognition performance. To assess this, we recruited nine additional participants from three age groups: young (9 years old in average), middle-aged (35 years old in average), and elderly (59 years old

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

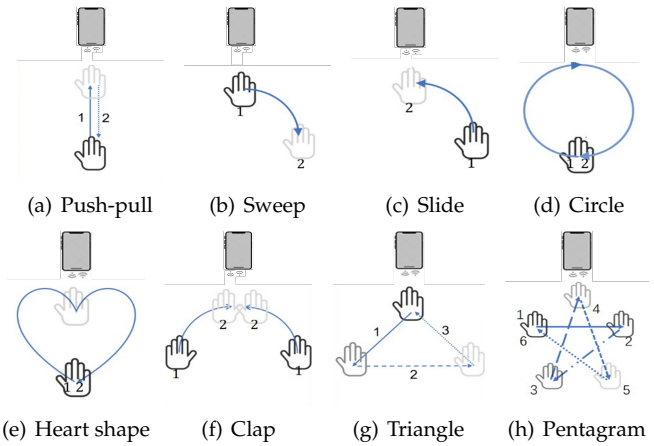


Fig. 25. Eight hand gestures.

in average). Each participant performed a gesture 20 times according to their natural habits. The experimental results, shown in Fig. 22, indicate that the average recognition accuracies in the 3-shot scenario were 93.05%, 97.45%, and 93.05% for the young, middle-aged, and elderly groups, respectively. These findings suggest that age does impact recognition performance, with middle-aged participants outperforming both younger and older groups. A potential explanation is that middle-aged individuals tend to have more refined gesture habits and greater control over their hand movements. In contrast, younger participants display less pronounced frequency shifts due to smaller-scale finger movements, while elderly participants perform gestures at a slower pace, leading to similarly subtle frequency shifts. Nevertheless, all participants achieved over 90% accuracy in the 3-shot scenario, demonstrating the system's robustness across age groups.

6.8 Different Gesture Datasets

To evaluate the generalization ability of Img2Acoustic, we conduct experiments on different gesture datasets, including 26 letter gestures and 8 hand gestures. Both datasets are collected under the default scenario. The 26 letter gestures are writing letters from 'A' to 'Z'. Fig. 25 shows the 8 hand gestures including actions of pushing, pulling, sweeping, sliding, drawing a pentagram, and *etc.*.

First, we evaluate a dataset containing 26 letter gestures, which is collected by 10 participants. In the letter gesture recognition task, we introduce top-1, top-3, and top-5 accuracy metrics. These metrics represent the model's ability to correctly predict the true label among the top-1, top-3, and top-5 predicted labels, respectively. The main goal is to simulate a function that suggests potential letter-writing options, which can facilitate human-computer interaction. As shown in Fig. 23, the top-1 and top-3 accuracies of letter gestures recognition are close to 85% and over 95% in the 3-shot case, respectively.

We then evaluate a dataset consisting of 8 hand gestures which is collected by 7 participants labeled by $G_1 \sim G_7$. These 8 hand gestures are selected as the testing dataset. They are not only based on frequent usage in human-computer interaction, but also are commonly employed in

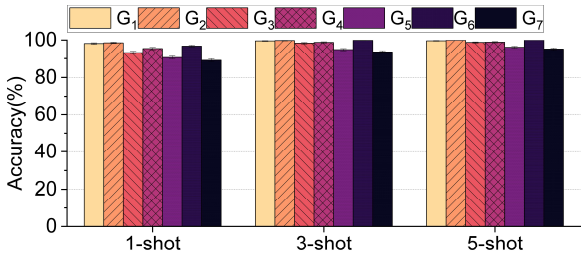


Fig. 26. The recognition accuracies of 8 hand gestures for different participants.

the evaluation of related works [51], [52]. Furthermore, these 8 gestures encompass the fundamental components of gesture actions, including movements away from and towards straight lines, and movements away from and towards curves. By utilizing these 8 hand gestures, we can verify the reliability of the system in terms of user-defined gestures. As shown in Fig. 23, the accuracy of hand gesture recognition is close to 95% in the 1-shot case. This is primarily due to the substantial disparities in the movements of these hand gestures. The average accuracy of different participants on the dataset of 8 hand gestures are shown in Fig. 26. We can observe that the average accuracy of all participants is always above 90% across different shot settings. In some cases, participants can achieve an average accuracy of 100% when providing 5 shots.

6.9 Running Time

For real-time interactive systems, response time is a key performance metric. We also assess the runtime performance of the proposed Img2Acoustic implemented on both server and smartphone platforms. Considering that the time of other modules in the APP is fixed, about 100 ms, we mainly measure the inference time of all Img2Acoustic models with data size of $(way, shot, 64, 19, 19)$ and $(1, 1, 3, 84, 84)$. The former represents support samples, and the latter represents a query sample. We assume that the support samples are subjected to feature extraction prior to being provided. The CPU configuration of the server is Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz, and the GPU configuration is NVIDIA RTX A6000. The CPU configuration of the mobile device is Qualcomm Snapdragon 865. We conduct tests per class from 2 to 20 classes by providing 5 shots. The results are shown in Fig. 27. Since we can not ensure that background processes do not start automatically, the data may fluctuate unexpectedly. It can be observed that the model's inference time is close to 200 ms when the way is 20 and users provide 5 shots. Obviously, the threshold for user experience is 200 ms. As a result, we recommend that the number of gesture categories entered by users should be no more than 10.

7 Discussion and Future Work

Although Img2Acoustic shows excellent potential in cross-modal acoustic gesture recognition, several challenges remain that need to be addressed to fully unlock its capabilities in HGR/HAR.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

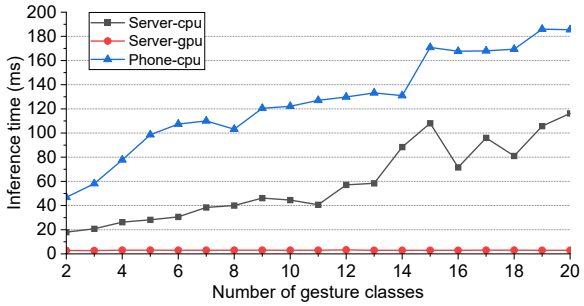


Fig. 27. The running time of Img2Acoustic with varying numbers of testing gestures.

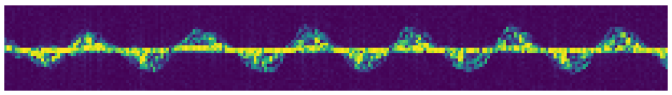


Fig. 28. Millimeterwave radar spectrogram.

7.1 Generalization to Other Modalities

In this work, we evaluate our proposed method with acoustic data. However, the method is not limited to acoustic data alone. It can be easily extended to other modalities, such as millimeter-wave Radar data. We show a segment of millimeter-wave radar data [20] in Fig. 28. The data exhibits noticeable frequency shifts, similar to the Doppler effect utilized in acoustic data. In the future, we aim to further enhance the performance of the model and explore its application to other modalities.

7.2 Improvement of Inference Speed

Currently, the average inference time of our system still takes hundreds of milliseconds, particularly when more categories are involved. In order to enhance the overall user experience, we have intentions to refine the model further and achieve swifter inference times. Considering that the current four-layer convolutional backbone network is already relatively straightforward, we plan to enhance the computational approach of the local matching network by reducing the number of vector multiplications, to further optimize the running time performance.

7.3 Robustness to Background Noise

We evaluate the system's robustness on different devices. The average recognition accuracy on Xiaomi device is around 90% when users provide 5 shots, but our system does not perform well with the mixture of background noise and useful signals. This implies that our model heavily relies on signal processing methods. In our future work, we plan to further improve the robustness of the model by introducing additional data augmentation techniques.

8 Conclusion

In this paper, we introduce Img2Acoustic, a novel cross-modal acoustic gesture recognition method. During training, we leverage open-source image datasets to train the model without requiring target modality data. Our approach incorporates TAAL, TALML, and local-global branch techniques to enhance task adaptation and improve domain transfer from source to target. Comprehensive evaluations show that Img2Acoustic effectively generalizes to various acoustic gesture recognition tasks, significantly reducing data collection costs and accelerating system deployment.

References

[1] M. T. I. Aumi, S. Gupta, M. Goel, E. Larson, and S. Patel, "Doplink: Using the doppler effect for multi-device interaction," in *Proceedings of ACM Ubicomp*, 2013, pp. 583–586.

[2] K.-Y. Chen, D. Ashbrook, M. Goel, S.-H. Lee, and S. Patel, "Airlink: sharing files between multiple devices using in-air gestures," in *Proceedings of ACM Ubicomp*, 2014, pp. 565–569.

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

[4] K. Ling, H. Dai, Y. Liu, A. X. Liu, W. Wang, and Q. Gu, "Ultragesture: Fine-grained gesture sensing and recognition," *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, pp. 2620–2636, 2020.

[5] 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 *Proceedings of ACM Ubicomp*, 2016, pp. 474–485.

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

[7] X. Wang, K. Sun, T. Zhao, W. Wang, and Q. Gu, "Dynamic speed warping: Similarity-based one-shot learning for device-free gesture signals," in *Proceedings of IEEE INFOCOM*. IEEE, 2020, pp. 556–565.

[8] Y. Jin, Y. Gao, Y. Zhu, W. Wang, J. Li, S. Choi, Z. Li, J. Chauhan, A. K. Dey, and Z. Jin, "Sonicasl: An acoustic-based sign language gesture recognizer using earphones," *Proceedings of ACM IMWUT*, vol. 5, no. 2, pp. 1–30, 2021.

[9] D. Li, J. Liu, S. I. Lee, and J. Xiong, "Room-scale hand gesture recognition using smart speakers," p. 462–475, 2022.

[10] Y. Wang, J. Shen, and Y. Zheng, "Push the limit of acoustic gesture recognition," *IEEE Transactions on Mobile Computing*, vol. 21, no. 5, pp. 1798–1811, 2020.

[11] M. Chen, P. Yang, J. Xiong, M. Zhang, Y. Lee, C. Xiang, and C. Tian, "Your table can be an input panel: Acoustic-based device-free interaction recognition," *Proceedings of ACM IMWUT*, vol. 3, no. 1, pp. 1–21, 2019.

[12] H. Du, P. Li, H. Zhou, W. Gong, G. Luo, and P. Yang, "Worddrecorder: Accurate acoustic-based handwriting recognition using deep learning," in *Proceedings of IEEE INFOCOM*. IEEE, 2018, pp. 1448–1456.

[13] M. Schrapel, M.-L. Stadler, and M. Rohs, "Pentelligence: Combining pen tip motion and writing sounds for handwritten digit recognition," in *Proceedings of ACM SIGCHI*, 2018, pp. 1–11.

[14] K. Wu, Q. Yang, B. Yuan, Y. Zou, R. Ruby, and M. Li, "Echowrite: An acoustic-based finger input system without training," *IEEE Transactions on Mobile Computing*, vol. 20, no. 5, pp. 1789–1803, 2020.

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

[16] H. Yin, A. Zhou, G. Su, B. Chen, L. Liu, and H. Ma, "Learning to recognize handwriting input with acoustic features," *Proceedings of ACM IMWUT*, vol. 4, no. 2, pp. 1–26, 2020.

[17] K. Ahuja, Y. Jiang, M. Goel, and C. Harrison, "Vid2doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition," in *Proceedings of ACM SIGCHI*, 2021, pp. 1–10.

[18] H. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Ploetz, "Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition," *Proceedings of ACM IMWUT*, vol. 4, no. 3, pp. 1–29, 2020.

[19] H. Kwon, B. Wang, G. D. Abowd, and T. Plötz, "Approaching the real-world: Supporting activity recognition training with virtual imu data," *Proceedings of ACM IMWUT*, vol. 5, no. 3, pp. 1–32, 2021.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3481443

[20] K. Deng, D. Zhao, Q. Han, Z. Zhang, S. Wang, A. Zhou, and H. Ma, "Midas: Generating mmwave radar data from videos for training pervasive and privacy-preserving human sensing tasks," *Proceedings of ACM IMWUT*, vol. 7, no. 1, pp. 1–26, 2023.

[21] S. Bhalla, M. Goel, and R. Khurana, "Imu2doppler: Cross-modal domain adaptation for doppler-based activity recognition using imu data," *Proceedings of ACM IMWUT*, vol. 5, no. 4, pp. 1–20, 2021.

[22] X. Wang, T. Liu, C. Feng, D. Fang, and X. Chen, "Rf-cm: Cross-modal framework for rf-enabled few-shot human activity recognition," *Proceedings of ACM IMWUT*, vol. 7, no. 1, pp. 1–28, 2023.

[23] D. Li, J. Liu, S. I. Lee, and J. Xiong, "Fm-track: pushing the limits of contactless multi-target tracking using acoustic signals," in *Proceedings of ACM Sensys*, 2020, pp. 150–163.

[24] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.

[25] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[26] J. Li, L. Huang, S. Shah, S. J. Jones, Y. Jin, D. Wang, A. Russell, S. Choi, Y. Gao, J. Yuan *et al.*, "Signring: Continuous american sign language recognition using imu rings and virtual imu data," *Proceedings of ACM IMWUT*, vol. 7, no. 3, pp. 1–29, 2023.

[27] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, pp. 4077–4078, 2017.

[28] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[29] T. Zhang and W. Huang, "Kernel relative-prototype spectral filtering for few-shot learning," in *Proceedings of IEEE/CVF ECCV*. Springer, 2022, pp. 541–557.

[30] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[31] S. Baik, M. Choi, J. Choi, H. Kim, and K. M. Lee, "Meta-learning with adaptive hyperparameters," *Advances in neural information processing systems*, vol. 33, pp. 20 755–20 765, 2020.

[32] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: a good embedding is all you need?" in *Proceedings of IEEE/CVF ECCV*. Springer, 2020, pp. 266–282.

[33] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, and V. N. Balasubramanian, "Charting the right manifold: Manifold mixup for few-shot learning," in *Proceedings of the IEEE/CVF WACV*, 2020, pp. 2218–2227.

[34] P. Li, S. Gong, C. Wang, and Y. Fu, "Ranking distance calibration for cross-domain few-shot learning," in *Proceedings of IEEE/CVF CVPR*, 2022, pp. 9099–9108.

[35] H. Wang and Z.-H. Deng, "Cross-domain few-shot classification via adversarial task augmentation," *arXiv preprint arXiv:2104.14385*, pp. 9099–9108, 2021.

[36] Y. Guo, N. C. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. Feris, "A broader study of cross-domain few-shot learning," in *Proceedings of IEEE/CVF ECCV*. Springer, 2020, pp. 124–141.

[37] H. Liang, Q. Zhang, P. Dai, and J. Lu, "Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder," in *Proceedings of IEEE/CVF ICCV*, 2021, pp. 9424–9434.

[38] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, "Cross-domain few-shot classification via learned feature-wise transformation," *arXiv preprint arXiv:2001.08735*, 2020.

[39] F. Zhou, P. Wang, L. Zhang, W. Wei, and Y. Zhang, "Revisiting prototypical network for cross domain few-shot learning," in *Proceedings of IEEE/CVF CVPR*, 2023, pp. 20 061–20 070.

[40] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of IEEE/CVF CVPR*, 2019, pp. 7260–7268.

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

[42] K. Sun, W. Wang, A. X. Liu, and H. Dai, "Depth aware finger tapping on virtual displays," in *Proceedings of ACM Mobisys*, 2018, pp. 283–295.

[43] N. Zhu, H. Chen, and Z. Yang, "Fine-grained multi-user device-free gesture tracking on today's smart speakers," in *Proceedings of IEEE MASS*. IEEE, 2021, pp. 99–107.

[44] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[45] D. A. Potvin, K. M. Parris, and R. A. Mulder, "Geographically pervasive effects of urban noise on frequency and syllable rate of songs and calls in silvereyes (zosterops lateralis)," *Proceedings of the Royal Society B: Biological Sciences*, vol. 278, no. 1717, pp. 2464–2469, 2011.

[46] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of IEEE/CVF CVPR*, 2017, pp. 2117–2125.

[47] J. Requeima, J. Gordon, J. Bronskill, S. Nowozin, and R. E. Turner, "Fast and flexible multi-task classification using conditional neural adaptive processes," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[48] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of IEEE/CVF ECCV*, 2018, pp. 3–19.

[49] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proceedings of IEEE/CVF CVPR, pages=1–8, year=2008, organization=IEEE*.

[50] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[51] R. Xiao, J. Liu, J. Han, and K. Ren, "Onefi: One-shot recognition for unseen gesture via cots wifi," in *Proceedings of ACM Sensys*, 2021, pp. 206–219.

[52] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi," *Proceedings of IEEE TPAMI*, vol. 44, no. 11, pp. 8671–8688, 2021.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on December 04,2024 at 04:21:29 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.