Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on AI for Good

Wearable Sensor-Based Few-Shot Continual Learning on Hand Gestures for Motor-Impaired Individuals via Latent Embedding Exploitation

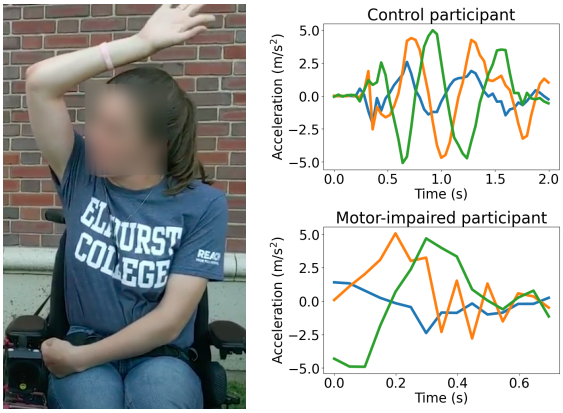
Riyad Bin Rafiq[1], **Weishi Shi**[1] and **Mark V. Albert**[1,2]
[1]Department of Computer Science and Engineering, University of North Texas
[2]Department of Biomedical Engineering, University of North Texas
riyadbinrafiq@my.unt.edu, {weishi.shi, mark.albert}@unt.edu

Abstract

Hand gestures can provide a natural means of human-computer interaction and enable people who cannot speak to communicate efficiently. Existing hand gesture recognition methods heavily depend on pre-defined gestures, however, motor-impaired individuals require new gestures tailored to each individual's gesture motion and style. Gesture samples collected from different persons have distribution shifts due to their health conditions, the severity of the disability, motion patterns of the arms, etc. In this paper, we introduce the Latent Embedding Exploitation (LEE) mechanism in our replay-based Few-Shot Continual Learning (FSCL) framework that significantly improves the performance of fine-tuning a model for out-of-distribution data. Our method produces a diversified latent feature space by leveraging a preserved latent embedding known as *gesture prior knowledge*, along with *intra-gesture divergence* derived from two additional embeddings. Thus, the model can capture latent statistical structure in highly variable gestures with limited samples. We conduct an experimental evaluation using the SmartWatch Gesture and the Motion Gesture datasets. The proposed method results in an average test accuracy of 57.0%, 64.6%, and 69.3% by using one, three, and five samples for six different gestures. Our method helps motor-impaired persons leverage wearable devices, and their unique styles of movement can be learned and applied in human-computer interaction and social communication. Code is available at: https://github.com/riyadRafiq/wearable-latent-embedding-exploitation.

1 Introduction

Hand gestures are a flexible and intuitive means of communication for human beings. With the advancement of wearable sensors and machine learning, gesture recognition has become quite popular for communication, smart home appliances, interactive entertainment, etc. [Rafiq *et al.*, 2023; Guo *et al.*, 2021]. Gesture-based interactions with wearables depend on specific presumptions about users' motor abilities. As a consequence, people with motor impairments face challenges in performing gestures with wearables that are widely adopted for the general public [Siean and Vatavu, 2021]. The severity of motor impairments leads to a different pattern of motion gestures and creates individual differences among the users [Vatavu and Ungurean, 2022]. It is social discrimination for this underrepresented population as it deprives them of completely leveraging those wearable devices. As the United Nations Sustainable Development Principle is *Leave no one behind*, increasing independence and including people with disabilities aid in achieving UN Sustainable Development Goals *Good health and well-being and Reduce inequalities* [Yu *et al.*, 2023].



(a) (b)

Figure 1: (a) An individual lacking fine motor skills performs hand gestures. (b) Sensor-based gesture samples of two different participants including a control participant (top) and a motor-impaired participant (bottom). Data samples are more variable and noisy for a motor-impaired individual rather than a control participant. Blue, orange, and green lines are acceleration values along the x, y, and z-axis respectively.

To tackle the problem, a large-scale labeled dataset is expected to build a robust hand gesture recognition method. However, this is impractical and cumbersome for motor-impaired individuals to participate in vast data collection. The transfer learning approach has been used to solve the problem. In transfer learning, a model is trained on a source domain and then fine-tuned to a target domain by

transferring knowledge from the prior learned task [Zhuang *et al.*, 2020]. But fine-tuning shows worse performance in a target domain with out-of-distribution samples [Kumar *et al.*, 2022]. Therefore, applying transfer learning alone cannot solve the problem, as the gesture data from the motor-impaired population are more variable and noisy than the control population (Figure 1), and limited data samples might not help the deep learning model capture the diverse patterns among each individual. To utilize limited training data, a unique approach has been proposed [Finn *et al.*, 2017] and in our case, few-shot transfer learning is an applicable solution.

In our case, another real-world problem is that all the gesture classes may not be available initially. For example, standard pre-defined gestures can be difficult to perform for individuals lacking fine motor skills. New unseen gestures may become accessible incrementally if motor-impaired individuals want to input their flexible and custom gestures. This context is referred to as a continual learning setting as the model involves learning a disjoint set of classes incrementally [Parisi *et al.*, 2019]. Continual learning posts two challenges, namely catastrophic forgetting [French, 1999] when the model's performance drops drastically on old classes and overfitting when the model is not capable of learning generalized features with a few training examples [Gidaris and Komodakis, 2018].

Many continual learning approaches including parameter regularization, functional regularization, replay strategy, etc. have become popular at present [Van de Ven and Tolias, 2019]. In this paper, we propose a novel method called Latent Embedding Exploitation (LEE) in our replay-based few-shot continual learning framework that can learn gesture classes incrementally from motor-impaired people. Specifically, in our framework, we utilize three latent embeddings from the feature extractor of a pre-trained model which is trained on the control subjects' gesture samples. The three embeddings are:

- a preserved latent embedding works as *gesture prior knowledge*,
- two additional latent embeddings known as temporary and learned embedding maintain a *intra-gesture divergence*.

They jointly aid a pre-trained model to be fine-tuned effectively with a few training examples from a motor-impaired individual. Ideally, the goal of LEE is to navigate the learned feature space toward a rich and diversified feature representation for variable and noisy data. Thus the fine-tuned model can capture the diverse pattern of unseen gesture classes with a few training examples. As a result, motor-impaired people can take full advantage of wearable devices with our proposed method. The major contributions of this paper are as follows:

- We explore wearable sensor-based hand gestures from the underrepresented population. In addition, we introduce the LEE mechanism in our replay-based few-shot continual learning framework that formulates the diverse gesture samples into a heterogeneous feature representation. Hence, the pre-trained model can be fine-tuned competently with a few training samples for each unseen class in a continual learning setup.
- We utilize two publicly available gesture datasets to demonstrate the performance of the proposed method. Our proposed method achieves competitive performance compared to existing methods.
- We experimentally show how latent embeddings can be leveraged to improve the performance of fine-tuning in the limited data of shifted distribution.

2 Related Work

A wide range of hand gesture recognition techniques has been explored by utilizing images and videos [Hu *et al.*, 2018; Zhou *et al.*, 2021], electromyography (EMG) [Caramiaux *et al.*, 2015] and wearable-sensors [Laput and Harrison, 2019; Kunwar *et al.*, 2022]. Among these techniques, vision-based approaches show poor performance due to complex backgrounds, varying light conditions, and the presence of another person in the background [Pisharady and Saerbeck, 2015; Mohamed *et al.*, 2021]. Moreover, high computational power is required to analyze high-quality video sequences and individuals might not be comfortable sharing live video streams due to privacy. On the contrary, sensors including accelerometers and gyroscopes are low-cost and widely available in current wearable devices. Therefore, in our work, we utilize wearable sensor-based motion data to solve the problem.

Prior work has been done using hand-crafted features such as mean, variance, median, maximum, minimum, etc. for classifying the hand gestures [Xie and Cao, 2016]. However, domain expertise is needed to prepare the necessary hand-crafted features which is time-consuming and the methods that utilize those features show poor performance in practice. On the contrary, deep learning approaches such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Transformer architectures have demonstrated significant performance in hand gesture classification with automatically extracted features from training data examples [Kunwar *et al.*, 2022; Nguyen-Trong *et al.*, 2021; Li *et al.*, 2019].

The goal of few-shot continual learning is to train new classes incrementally with few data instances. In order to tackle few-shot learning problem, metric-learning [Kaya and Bilge, 2019] , meta-learning [Finn *et al.*, 2017] and multi-task learning [Zhang and Yang, 2021] have been proposed. In the field of hand gesture recognition, camera data [Wu *et al.*, 2012; Stewart *et al.*, 2020] and EMG signals [Rahimian *et al.*, 2021] have been used for few-shot learning. However, Xu et al. [2022] proposed a hand gesture customization framework that can learn novel hand gesture classes incrementally with a few training examples. Kimura [2022] also proposed a self-supervised method for few-shot hand gesture recognition using wearable sensor data. However, they used only control participants' data. In addition, these methods would not work for motor-impaired individuals as the data instances are varying and noisy. Although Malu et al. [2018] and Kim et al. [2019] explored the smartwatch interactions for people with upper body motor impairments, they experimented

with the touch gestures only. The primary difference between our approach and existing works is that they did not consider the out-of-distribution samples for motion gestures. Directly fine-tuning a pre-trained model may cause a sudden performance decay. The objective of our method is to adapt a model that generates an enhanced feature representation via *gesture prior knowledge* exploitation and *intra-gesture divergence* exploration to incrementally learn novel gestures with few training examples from motor-impaired individuals.

3 Proposed Method

3.1 Problem Statement

A domain is defined as a joint probability distribution $\mathbb{P}_{x,y}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the instance space and label space, respectively [Ding and Fu, 2017; Qian *et al.*, 2021]. In our setting, we have two domains including source domain, $\mathcal{D}^s = \{(x_i, y_i)\}_{i=1}^{n_s}$ and target domain, $\mathcal{D}^t = \{(x_i, y_i)\}_{i=1}^{n_t}$ where $n_t << n_s$. Each sample, $x \in \mathbb{R}^{L\times 3}$ denotes a signal of L length with three-axis motion values collected from wearable sensors at each timestamp. The two domains have the same feature space ($\mathcal{X}^s = \mathcal{X}^t$) but different label spaces ($\mathcal{Y}^s \neq \mathcal{Y}^t$). In addition to it, they have different probability distributions i.e. $P^s(x_i, y_i) \neq P^t(x_i, y_i)$. The data distribution for $\mathcal{D}^t$ is harder to learn than $\mathcal{D}^s$ due to high-level noise e.g. $\mathrm{H}[P^t(x)] >> \mathrm{H}[P^s(x)]$ where H denotes the entropy. Therefore, any statistical learner needs more samples from $\mathcal{D}^t$ to converge to achieve the same level of accuracy as models trained on $\mathcal{D}^s$. In $\mathcal{D}^t$, the classes $\mathcal{C}_1$, $\mathcal{C}_2$, ..., $\mathcal{C}_n$ are incrementally accessible at the training time. While accessing a new class, $\mathcal{C}_i$, a memory buffer stores training examples from old classes such as $\mathcal{C}_1$, $\mathcal{C}_2$, ..., $\mathcal{C}_{i-1}$. Motor-impaired individuals may face challenges in providing many consistent data samples. As a result, it becomes more difficult to train the model with such limited samples. The goal is to build a pre-trained model in the source domain $f : \mathcal{X}^s \rightarrow \mathcal{Y}^s$ that can be fine-tuned to learn currently available classes from the target domain with few training examples. We design the method in such a way that should reduce the memory usage compared to traditional replay buffers.

3.2 Overall Framework

Sensor readings from motor-impaired individuals vary substantially due to factors such as health conditions, severity of disability, movement patterns of arms, etc. (Figure 1b). Therefore, it is critical for a pre-trained model to adapt to such data samples from sensor readings. Machine learning models without considering out-of-distribution data often result in large performance degradation. For example, a model trained on the data from control participants often fails to capture unique patterns and performs poorly on specific populations such as Parkinson's patients [Bin Rafiq *et al.*, 2020]. It is difficult to collect a large volume of diverse labeled data from motor-impaired individuals. Moreover, individuals may need their own, custom flexible gestures from time to time in human-computer interaction and social communication.

In this paper, we propose a novel technique where the model learns new gesture classes incrementally by utilizing multifunctional latent embeddings. Our method considers three latent embeddings instead of a single representation compared to Autoencoders [Bank *et al.*, 2020]. As shown in Figure 2, a latent embedding from the control population is preserved by leveraging the feature extractor (deep encoder) of a pre-trained model. This preserved latent embedding works as *gesture prior knowledge* to assist the model to incrementally learn unseen out-of-distribution gesture samples and prevent overfitting. Two additional identical feature extractors are utilized to produce two latent embeddings with available gesture classes from the motor-impaired subjects, and one of them is being updated during training. As a consequence, the learned latent embedding has a strong capability in classifying highly variable data during inference. The total loss of the model in weighted summation is as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{ci} + \beta\mathcal{L}_{ii} + \mathcal{L}_{cls} \tag{1}$$

where $\mathcal{L}_{ci}$ is the loss of discrimination between preserved and learned embedding, $\mathcal{L}_{ii}$ is the loss of discrimination between temporary and learned embedding and $\mathcal{L}_{cls}$ is the classification loss. α and β are trade-off hyper-parameters where $\alpha + \beta = 1$. While minimizing the loss in the training stage, the model exhibits complementary learning behavior by adaptively adjusting its focus between exploitative and explorative representation learning.

3.3 Complementary Learning Paradigm

The complementary learning system plays an important role in the human brain where the hippocampus and the neocortex function in a complementary manner to learn complex behavior [Perrusquía, 2022; Blakeman and Mareschal, 2020]. Our learning strategy for new classes is inspired by this *complementary learning paradigm*. The representation space holds the same classes closer under the effect of classification objective while training a deep learning model. However, the model struggles to learn a robust latent space with fewer training examples from out-of-distribution data [Yang *et al.*, 2021]. Therefore, we model the representation space generation process by utilizing both the classification and the embedding discrimination objectives. In our method, we introduce three multipurpose latent embeddings including preserved control embedding, temporary sample embedding, and learned embedding. The preserved control embedding contains the representation space of the expected pattern of gestures from the source domain. The learned embedding constructs a latent statistical structure with the help of preserved and temporary embedding. Thus, the constructed latent space sufficiently narrows the features of the same-class data with limited training examples.

Latent Embedding Exploitation

The learned latent embedding exploits the preserved latent embedding (*gesture prior knowledge*) to enlarge and diversify the feature space. The network architecture is expected to increase the similarity of the feature space between the preserved and the learned embedding. We utilize the feature extractor, f_θ (deep encoder), from the pre-trained model to preserve a latent embedding, $\mathbf{z_c} = f_\theta(\mathcal{X}^s)$ where $\mathcal{X}^s$ is the feature space from the source domain i.e. control participants.

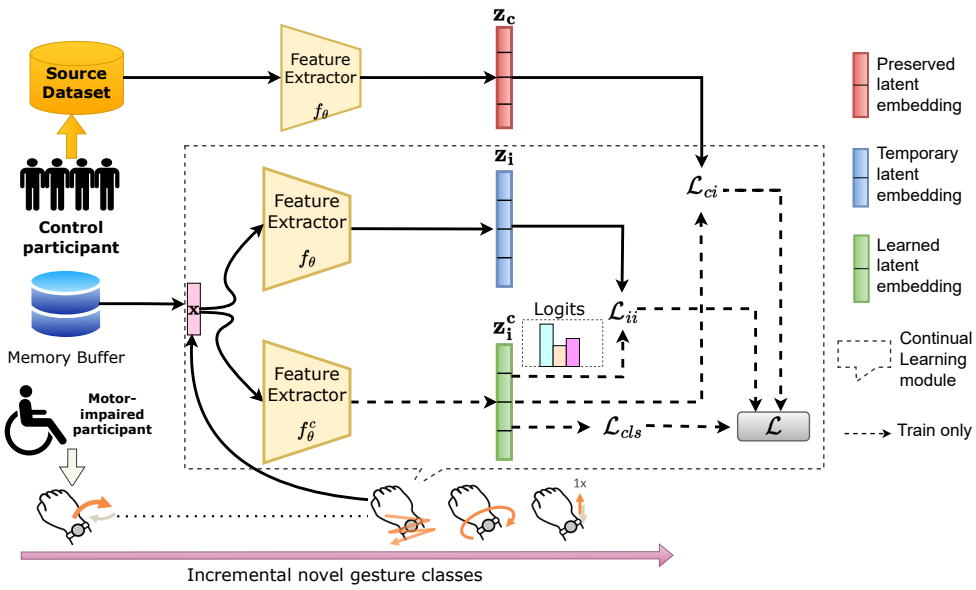


Figure 2: The complete framework containing the LEE mechanism. A latent embedding, $\mathbf{z_c}$ from the control population is preserved to work as *gesture prior knowledge*. In addition to it, two latent embeddings, $\mathbf{z_i}$ and $\mathbf{z_i^c}$ function to maintain *intra-gesture divergence*. The memory buffer saves the training samples from old gesture classes and provides them while training on a novel class.

The input of our model is denoted as $\mathbf{x}$. In our continual learning module, two additional identical feature extractors, f_θ and f_θ^c draw out temporary latent embedding, $\mathbf{z_i} = f_\theta(\mathbf{x})$ and learned latent embedding, $\mathbf{z_i^c} = f_\theta^c(\mathbf{x})$ respectively. To this end, we require to expand the similarity between $\mathbf{z_c}$ and $\mathbf{z_i^c}$ as the following loss:

$$\mathcal{L}_{ci}(f_\theta^c; \mathbf{x}) = 1 - \mathcal{S}_c(\mathbf{z_c}, \mathbf{z_i^c}) \tag{2}$$

where $\mathcal{S}_c$ is the cosine similarity between $\mathbf{z_c}$ and $\mathbf{z_i^c}$ and it can be defined as follows:

$$\mathcal{S}_c(\mathbf{z_c}, \mathbf{z_i^c}) = \frac{\mathbf{z_c} \cdot \mathbf{z_i^c}}{||\mathbf{z_c}||||\mathbf{z_i^c}||} \tag{3}$$

Latent Embedding Exploration

Simultaneously, the learned latent embedding aims to maximize the distance from the identical temporary sample embedding throughout the training which is called *intra-gesture divergence*. This action works as a way of exploration for a wide feature space. As a result, the learned latent embedding captures tailored and generalizable feature representation to learn novel gesture classes. To minimize the similarity between $\mathbf{z_i}$ and $\mathbf{z_i^c}$ is identical as follows:

$$\mathcal{L}_{ii}(f_\theta, f_\theta^c; \mathbf{x}) = \mathcal{S}_c(\mathbf{z_i}, \mathbf{z_i^c}) \tag{4}$$

Learning Objective

The learning objective of the model is to identify the gesture classes which is a transformation of the input sensor signals to a gesture category. Therefore, we utilize class labels in the final classification layer to guide the learned latent embedding during the training stage. We adopt standard cross-entropy loss for the classification task:

$$\mathcal{L}_{cls}(f_\theta^c; \mathcal{X}^t, \mathcal{Y}^t) = -\mathbb{E}_{(x,y)\in\mathcal{X}^t\times\mathcal{Y}^t} \sum_{c=1}^{C} y \log \delta_c(f_\theta^c(\mathbf{x})) \tag{5}$$

where C represents the number of classes, y is the true gesture label, $\delta_c(f_\theta^c(\mathbf{x}))$ is the predicted probability and δ_c is the softmax function.

4 Experiments

4.1 Datasets

The SmartWatch Gesture Dataset [Porzi *et al.*, 2013; Costante *et al.*, 2014] was built for interacting with mobile applications using arm gestures. This dataset contains 20 distinct gestures from eight different subjects. A first-generation Sony smartwatch with a built-in 3-axis accelerometer was worn on the user's right wrist while performing 20 repetitions for each gesture. In total, 3200 sequences were collected and each sequence contains 3-axis acceleration data. We use this dataset as our source domain to build the pre-trained model.

The Motion Gesture Dataset [Vatavu and Ungurean, 2022] was built to understand the gesture articulation of people with upper-body motor impairments. Six different motion gestures were collected by a group of 12 people (six male and six female) with upper-body motor impairments, ranging ages from 27 to 65 years. The participants had a wide range of disabilities including Spinal cord injury, Traumatic brain injury, Multiple sclerosis, Parkinson's disease, etc. A Samsung Gear Fit 2 smartwatch was used by the participants to collect the wrist gesture's accelerometer data. Each participant repeated each gesture eight times. We utilize this dataset in our few-shot continual learning setting.

4.2 Implementation Details

The sequence length of the data samples varies extensively. As a result, we apply a linear interpolation technique to our source and target datasets so that the sequence length ($L = 50$) of each data sample is constant throughout the

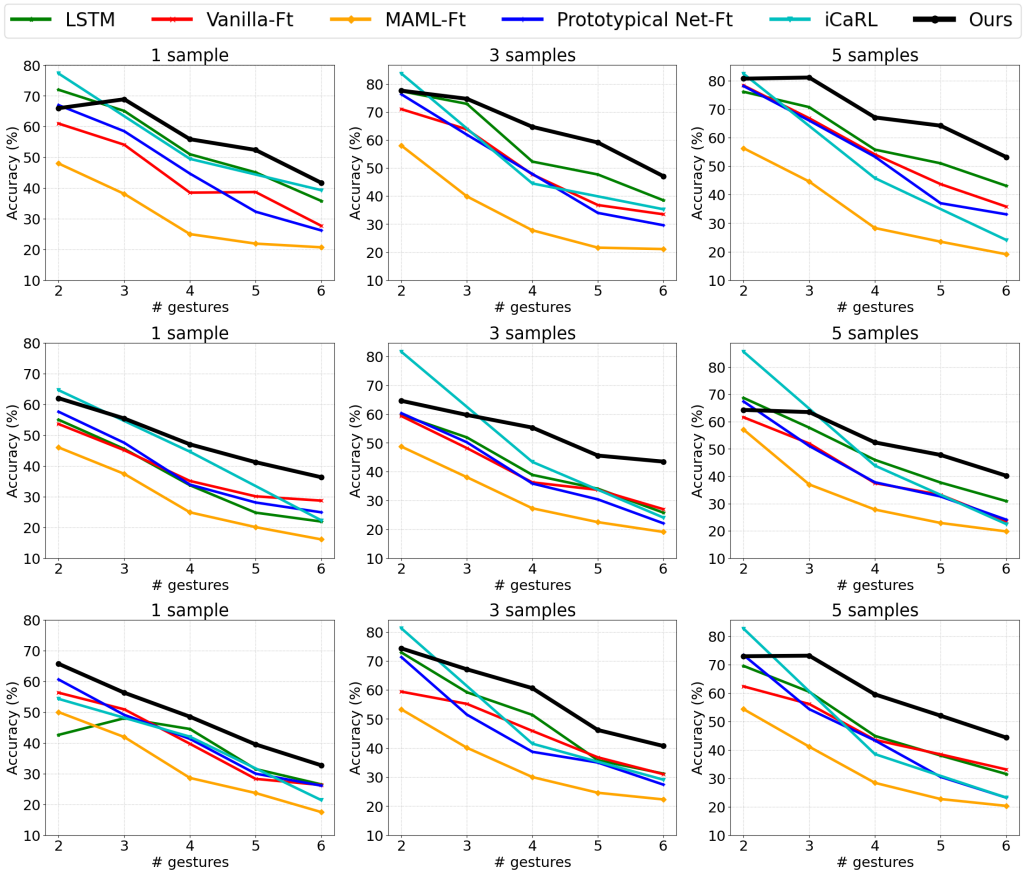


Figure 3: Test accuracies for a motor-impaired individual with Spinal cord injury (top row), an individual with Parkinson's disease (middle row), and a participant with Multiple sclerosis (bottom row) in a few-shot continual learning setting. The accuracy represents the total accuracy over all the gesture classes encountered trained with one, three, and five samples.

experiments. While working with the sensor data, outlier features can negatively influence the results. Therefore, dataset standardization is conducted by removing the mean and scaling it according to the interquartile range for each feature. We select 16 out of 20 gestures from the source domain to build a pre-trained model because we do not want any overlapped gestures between the source domain and the target domain. We follow the leave-one-subject-out strategy to pre-train the model. Throughout all experiments, we used the same subjects for a fair comparison. For our architecture, the feature extractor contains one LSTM layer with 64-dimensional hidden representation, one fully connected layer with 14 units, and one dropout layer with a value of 0.5 between them. A fully connected layer is used as the classification layer. The network architecture remains the same throughout all experiments. The Adam optimizer with learning rate 10^{-3} is used for the few-shot setup. In the few-shot continual learning setting, since each gesture class contains very few training examples, the epoch is set to 15 and the mini-batch contains all examples. We run each experiment 10 times with five different orders of the gesture classes. We report the average accuracy with one, three, and five training examples over all the encountered gesture classes. As this is a continual learning setup, we also report the class-wise macro F1 score and *forgetting* metric to understand each gesture's performance individually. Our memory buffer stores $60\times$ fewer examples compared to traditional replay buffers [Rebuffi *et al.*, 2017].

4.3 Baselines and Compared Methods

We compare our methods with different closely related approaches in the few-shot continual learning setting. Since our method utilizes LSTM in the network architecture, we consider comparing it with an LSTM classifier that learns the gesture classes incrementally with a few training examples. Vanilla-Ft, MAML-Ft [Finn *et al.*, 2017], and Prototypical Net-Ft [Snell *et al.*, 2017] involve fine-tuning the pre-trained models on the few-shot classes. We also compare our method with iCaRL [Rebuffi *et al.*, 2017]. We assume that the memory buffer exists in all methods for a fair comparison.

4.4 Experimental Results

We compare the average accuracy of the proposed method with other approaches. Figure 3 shows the test accuracies for three participants including a motor-impaired individual with Spinal cord injury, an individual with Parkinson's disease, and an individual with Multiple sclerosis. In most cases, our LEE method outperforms other techniques. We observe that the iCaRL classifier occasionally shows better accuracy than our method while learning two initial gestures. But

Gesture classes	Methods					
	LSTM	Vanilla-Ft	MAML-Ft	Prototypical Net-Ft	iCaRL	**Ours-LEE**
Gesture 1	$\underline{0.44 \pm 0.08}$	0.31 ± 0.07	0.13 ± 0.10	0.24 ± 0.09	0.01 ± 0.02	$\mathbf{0.53 \pm 0.19}$
Gesture 2	$\underline{0.21 \pm 0.11}$	0.16 ± 0.05	0.12 ± 0.08	0.10 ± 0.04	0.04 ± 0.08	$\mathbf{0.28 \pm 0.13}$
Gesture 3	0.35 ± 0.15	$\underline{0.40 \pm 0.07}$	0.12 ± 0.07	0.32 ± 0.05	0.16 ± 0.15	$\mathbf{0.58 \pm 0.17}$
Gesture 4	$\underline{0.34 \pm 0.17}$	0.28 ± 0.10	0.13 ± 0.08	0.22 ± 0.11	0.01 ± 0.02	$\mathbf{0.52 \pm 0.20}$
Gesture 5	$\underline{0.40 \pm 0.18}$	0.28 ± 0.19	0.15 ± 0.03	0.21 ± 0.12	0.30 ± 0.17	$\mathbf{0.53 \pm 0.24}$
Gesture 6	0.29 ± 0.14	0.28 ± 0.14	0.10 ± 0.13	0.21 ± 0.13	$\underline{0.35 \pm 0.15}$	$\mathbf{0.43 \pm 0.14}$

Table 1: Gesture class-wise average macro F1 score for a motor-impaired individual with Spinal cord injury in a few-shot continual learning setting (mean±std). We report the scores after six gestures are trained with five training examples. The best macro F1 score is highlighted in bold whereas the second-best score is underlined.

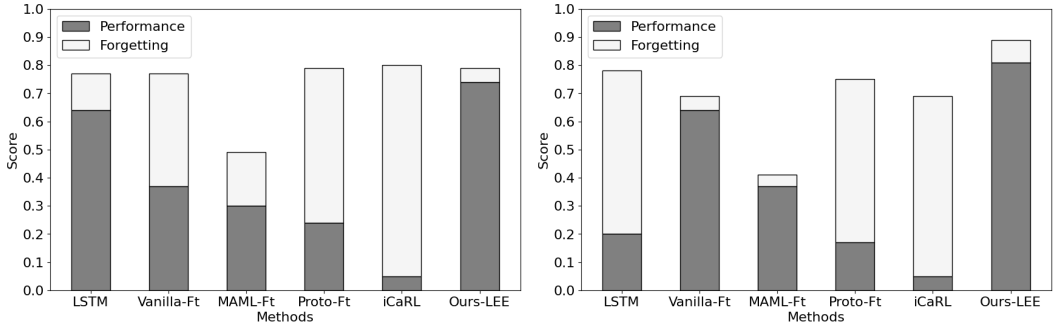


Figure 4: Performance-*forgetting* scaled score for *Gesture 1* (left) and *Gesture 3* (right) for a motor-impaired individual with Spinal cord injury after six gestures are trained with five training examples.

for the rest of the incrementally added classes, LEE always performs better than the iCaRL classifier. The accuracy of the iCaRL classifier significantly drops for new gesture classes because it fails to capture the diverse and highly variable pattern of unseen gestures with few training examples. The performance increases with the sample size for all methods. Surprisingly, the fine-tuning approach fails compared to the basic LSTM classifier and our LEE.

In a continual learning setup, it is important to perform well in old classes while trained on a new class. Therefore, in Table 1, we report the class-wise F1 scores after six gestures are trained with five training examples. Our method always provides a higher macro F1 score than other methods. LEE has a 12.3% higher F1 score for all gesture classes. Figure 4 shows performance and *forgetting* score for *Gesture 1* and *Gesture 3* after six gestures are trained with five samples. Our LEE method has better performance with less forgetting.

5 Ablation Study

5.1 Loss Hyperparameter Sensitivity Analysis

We report the effect of loss hyperparameter sensitivity. We focus only on α as it complements the other hyperparameter (β) in our continual learning module. We choose α to exploit the *gesture prior knowledge*, selected from $\alpha \in \{0.01, 0.05, 0.5, 0.1, 0.9\}$. According to Figure 5 (left), LEE provides robust accuracy with a wide range of hyperparameters after learning six gestures using five training examples.

5.2 Number of Participants and Gestures in Source Domain

We experiment to produce the preserved latent embedding with a different number of participants from the source domain. We achieve higher accuracy with one and three samples using seven different participants (Figure 5 middle). Apart from this, the proposed method is invariant to the number of participants from the source domain. However, it is preferable to utilize a large number of control participants in the source domain to capture a more diversified representation space. We also conduct experiments with a different number of gestures from the source domain to generate the preserved latent embedding (Figure 5 right). Though we get the highest accuracy using 16 gestures, we observe that the accuracy does not change for other different numbers of gestures. Therefore, the observation illustrates that our method is robust to the number of gesture classes.

5.3 Significance of Embeddings

We conduct experiments to investigate how the preserved latent embedding and the temporary latent embedding contribute to our proposed method. We apply LEE without one of those embeddings, one at a time, and evaluate the performance. Figure 6 (left) and (middle) show that removing either component of interest results in a less tridiagonal-shaped confusion matrix, indicating a drop in model performance and robustness. We further confirm from Figure 6 (right) that both embeddings jointly contribute to robust performance, and thus these embeddings are the foundations for learning unseen gestures incrementally with limited training examples.

6 Use Cases and Social Impact

Wearable sensor-based gesture recognition is becoming popular in many areas including communication, controlling home appliances, and interactive entertainment. As most

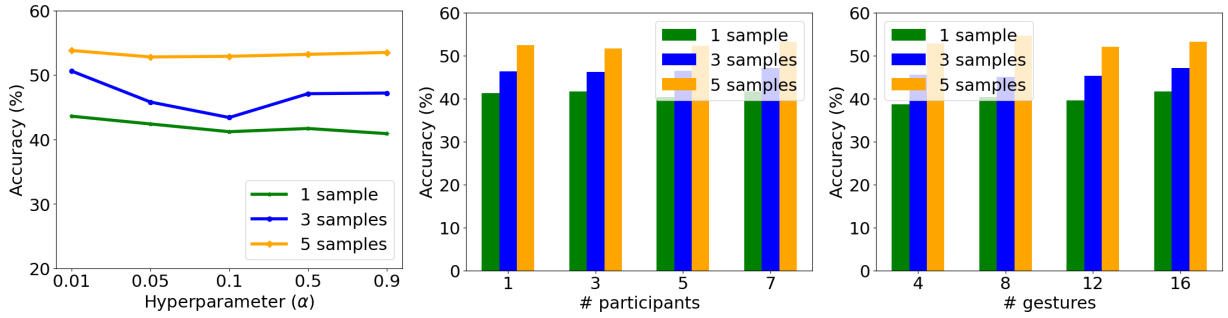


Figure 5: Accuracy for different hyperparameter values (left), number of participants (middle), and number of gestures (right) from the source domain when the preserved latent embedding is produced. We report the accuracy after six gestures are trained with five training examples.

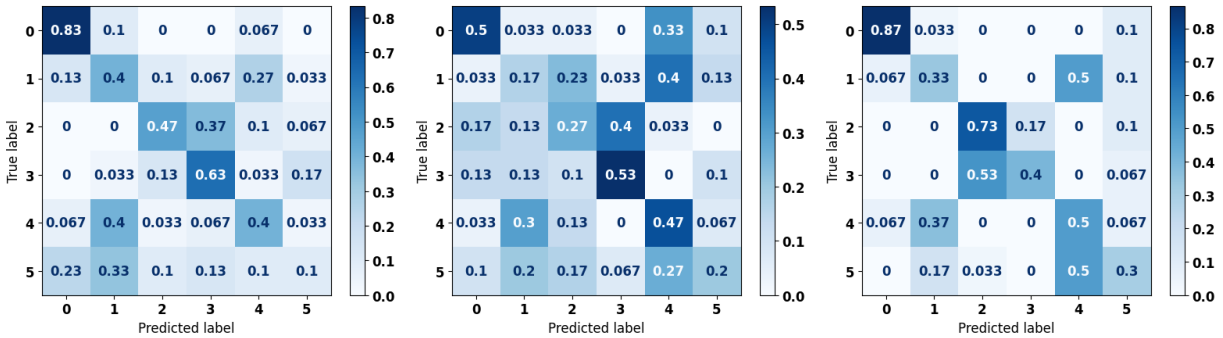


Figure 6: The confusion matrices for six gesture classes after training with five training examples. left: w/o preserved latent embedding ($\alpha = 0$); middle: w/o temporary latent embedding ($\alpha = 1$); right: LEE ($\alpha = 0.5$).

research doesn't include the motor-impaired population, those individuals face challenges using wearable devices for gesture recognition and communication. We believe our contribution can be impactful for those who need more than a set of pre-defined gestures. Gesture recognition exists for standard movements such as sign language for speech-impaired people but sign language can be difficult to perform for individuals lacking fine motor skills. Our work is part of a fast and flexible gesture-to-speech recognition system that we are developing in collaboration with Shirley Ryan AbilityLab [1]. The need for such solutions is underscored by the prevalence of motor impairments that also impact speech. 12.1% of the population has a motor disability [CDC, 2023] while 7.6% have a speech disorder [NIDCD, 2024] through Stroke ($795,000$ cases each year), Parkinson's disease (1 million), Multiple sclerosis ($727,000$), Spinal cord injury ($294,000$) and Cerebral palsy ($764,000$) [Wallin *et al.*, 2019; White and Black, 2016]. The proposed method has the potential to transform the lives of these individuals by providing a more natural and efficient mode of communication, improving quality of life, enhancing interaction, and reducing the burden on caregivers.

7 Conclusion and Future Work

Hand gestures are natural and flexible means of communication. Available wearable sensor-based hand gesture solutions are widely adopted for the normal population and these solutions fail to capture highly variable and inconsistent data samples from motor-impaired people. Moreover, in the real world, motor-impaired individuals face challenges in performing predefined gestures, and many valuable use cases rely on acquiring new gestures. However, a substantial amount of data samples is needed to develop a strong hand gesture recognition method. Therefore, we introduce a novel method called Latent Embedding Exploitation (LEE) to learn novel gesture classes incrementally using a few samples from motor-impaired individuals. We experimentally show that our method outperforms the existing baselines. Our method helps motor-impaired persons leverage wearable devices and their unique movement styles can be learned and applied in human-computer interaction and social communication. By enabling meaningful interactions with motor-impaired individuals and seamlessly integrating wearable devices into their daily lives, we open the gate to collecting invaluable data from this underrepresented group in real-world scenarios. This data collection paradigm can play a central role in facilitating the advancements in other machine learning research, benefiting not only motor-impaired individuals but also contributing to broader technological innovation. In the future, we will integrate our method with a wearable application and online learning will be explored to assist motor-impaired individuals to input their custom, flexible gestures in real-time. Furthermore, we will survey to collect the opinions of the population and tailor our approach accordingly.

[1] https://www.sralab.org/

Acknowledgments

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD101899.

References

[Bank *et al.*, 2020] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv preprint arXiv:2003.05991*, 2020.

[Bin Rafiq *et al.*, 2020] Riyad Bin Rafiq, Francois Modave, Shion Guha, and Mark V Albert. Validation methods to promote real-world applicability of machine learning in medicine. In *2020 3rd International Conference on Digital Medicine and Image Processing*, pages 13–19, 2020.

[Blakeman and Mareschal, 2020] Sam Blakeman and Denis Mareschal. A complementary learning systems approach to temporal difference learning. *Neural Networks*, 122:218–230, 2020.

[Caramiaux *et al.*, 2015] Baptiste Caramiaux, Marco Donnarumma, and Atau Tanaka. Understanding gesture expressivity through muscle sensing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(6):1–26, 2015.

[CDC, 2023] CDC. Disability impacts all of us, 2023. Accessed: May 13, 2024.

[Costante *et al.*, 2014] Gabriele Costante, Lorenzo Porzi, Oswald Lanz, Paolo Valigi, and Elisa Ricci. Personalizing a smartwatch-based gesture interface with transfer learning. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 2530–2534. IEEE, 2014.

[Ding and Fu, 2017] Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27(1):304–313, 2017.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[French, 1999] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

[Gidaris and Komodakis, 2018] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.

[Guo *et al.*, 2021] Lin Guo, Zongxing Lu, and Ligang Yao. Human-machine interaction sensing technology based on hand gesture recognition: A review. *IEEE Transactions on Human-Machine Systems*, 51(4):300–309, 2021.

[Hu *et al.*, 2018] Ting-Kuei Hu, Yen-Yu Lin, and Pi-Cheng Hsiu. Learning adaptive hidden layers for mobile gesture recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[Kaya and Bilge, 2019] Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019.

[Kim *et al.*, 2019] Jee-Eun Kim, Masahiro Bessho, and Ken Sakamura. Towards a smartwatch application to assist students with disabilities in an iot-enabled campus. In *2019 IEEE 1st Global Conference on Life Sciences and Technologies (LifeTech)*, pages 243–246. IEEE, 2019.

[Kimura, 2022] Naoki Kimura. Self-supervised approach for few-shot hand gesture recognition. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–4, 2022.

[Kumar *et al.*, 2022] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[Kunwar *et al.*, 2022] Utkarsh Kunwar, Sheetal Borar, Moritz Berghofer, Julia Kylmälä, Ilhan Aslan, Luis A Leiva, and Antti Oulasvirta. Robust and deployable gesture recognition for smartwatches. In *27th International Conference on Intelligent User Interfaces*, pages 277–291, 2022.

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

[Li *et al.*, 2019] Chenyang Li, Xin Zhang, Lufan Liao, Lianwen Jin, and Weixin Yang. Skeleton-based gesture recognition using several fully connected layers with path signature features and temporal transformer module. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8585–8593, 2019.

[Malu *et al.*, 2018] Meethu Malu, Pramod Chundury, and Leah Findlater. Exploring accessible smartwatch interactions for people with upper body motor impairments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[Mohamed *et al.*, 2021] Noraini Mohamed, Mumtaz Begum Mustafa, and Nazean Jomhari. A review of the hand gesture recognition system: Current progress and future directions. *IEEE Access*, 9:157422–157436, 2021.

[Nguyen-Trong *et al.*, 2021] Khanh Nguyen-Trong, Hoai Nam Vu, Ngon Nguyen Trung, and Cuong Pham. Gesture recognition using wearable sensors with bi-long short-term memory convolutional neural networks. *IEEE Sensors Journal*, 21(13):15065–15079, 2021.

[NIDCD, 2024] NIDCD. Quick statistics about voice, speech, and language, 2024. Accessed: May 13, 2024.

[Parisi *et al.*, 2019] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

[Perrusquía, 2022] Adolfo Perrusquía. Human-behavior learning: A new complementary learning perspective for optimal decision making controllers. *Neurocomputing*, 489:157–166, 2022.

[Pisharady and Saerbeck, 2015] Pramod Kumar Pisharady and Martin Saerbeck. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141:152–165, 2015.

[Porzi *et al.*, 2013] Lorenzo Porzi, Stefano Messelodi, Carla Mara Modena, and Elisa Ricci. A smart watch-based gesture recognition system for assisting people with visual impairments. In *Proceedings of the 3rd ACM international workshop on Interactive multimedia on mobile & portable devices*, pages 19–24, 2013.

[Qian *et al.*, 2021] Hangwei Qian, Sinno Jialin Pan, and Chunyan Miao. Latent independent excitation for generalizable sensor-based cross-person activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11921–11929, 2021.

[Rafiq *et al.*, 2023] Riyad Bin Rafiq, Syed Araib Karim, and Mark V Albert. An lstm-based gesture-to-speech recognition system. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 430–438. IEEE, 2023.

[Rahimian *et al.*, 2021] Elahe Rahimian, Soheil Zabihi, Amir Asif, S Farokh Atashzar, and Arash Mohammadi. Few-shot learning for decoding surface electromyography for hand gesture recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1300–1304. IEEE, 2021.

[Rebuffi *et al.*, 2017] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[Siean and Vatavu, 2021] Alexandru-Ionut Siean and Radu-Daniel Vatavu. Wearable interactions for users with motor impairments: systematic review, inventory, and research implications. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–15, 2021.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[Stewart *et al.*, 2020] Kenneth Stewart, Garrick Orchard, Sumit Bam Shrestha, and Emre Neftci. Online few-shot gesture learning on a neuromorphic processor. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 10(4):512–521, 2020.

[Van de Ven and Tolias, 2019] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

[Vatavu and Ungurean, 2022] Radu-Daniel Vatavu and Ovidiu-Ciprian Ungurean. Understanding gesture input articulation with upper-body wearables for users with upper-body motor impairments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022.

[Wallin *et al.*, 2019] Mitchell T Wallin, William J Culpepper, Jonathan D Campbell, Lorene M Nelson, Annette Langer-Gould, Ruth Ann Marrie, Gary R Cutter, Wendy E Kaye, Laurie Wagner, Helen Tremlett, et al. The prevalence of ms in the united states: a population-based estimate using health claims data. *Neurology*, 92(10):e1029–e1040, 2019.

[White and Black, 2016] Non-Hispanic White and Non-Hispanic Black. Spinal cord injury (sci) facts and figures at a glance. 2016.

[Wu *et al.*, 2012] Di Wu, Fan Zhu, and Ling Shao. One shot learning gesture recognition from rgbd images. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 7–12. IEEE, 2012.

[Xie and Cao, 2016] Renqiang Xie and Juncheng Cao. Accelerometer-based hand gesture recognition by neural network and similarity matching. *IEEE Sensors Journal*, 16(11):4537–4545, 2016.

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

[Yang *et al.*, 2021] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395*, 2021.

[Yu *et al.*, 2023] Lu Yu, Malvina Nikandrou, Jiali Jin, and Verena Reiser. Quality-agnostic image captioning to safely assist people with vision impairment. *arXiv preprint arXiv:2304.14623*, 2023.

[Zhang and Yang, 2021] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[Zhou *et al.*, 2021] Benjia Zhou, Yunan Li, and Jun Wan. Regional attention with architecture-rebuilt 3d network for rgb-d gesture recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3563–3571, 2021.

[Zhuang *et al.*, 2020] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.