Hand Gesture Recognition With Flexible Capacitive Wristband Using Triplet Network in Inter-Day Applications

Tiantong Wang®, Student Member, IEEE, Yunbiao Zhao, Member, IEEE, and Qining Wang®, Senior Member, IEEE

Abstract—Human-machine interfaces for hand gesture recognition across multiple sessions and days of doffing and re-donning while maintaining acceptable recognition accuracy are still challenging. In this paper, a flexible wristband, which was integrated with a highly sensitive capacitive pressure sensing array, was used for inter-day hand gesture recognition. The performance of the entire system was further improved by utilizing a triplet network for deep feature embedding. Seven hand gestures were included into the gesture set, and inter-day experiments which lasted for five consecutive days with three sessions on each day were conducted. Five healthy subjects participated in the experiment. Between each session, the wristband was doffed, and re-donned before the next session. The triplet network achieved an average recognition accuracy of 91.98% across all the sessions of all the subjects, and yielded a higher classification result ($p < 0.05$) over the convolutional neural network trained with softmax-cross-entropy loss (with an average accuracy of 84.65%). Furthermore, we also found that the capacitive array size had an evident influence on the inter-day classification result. The array with the full size (thirty-two channels) achieved a higher average recognition accuracy over all the down-sampled arrays. This work demonstrated the feasibility of improving the hand gesture recognition performance over days of usage by fabricating a wearable, flexible multi-channel capacitive wristband and implementing the triplet network.

Index Terms—Capacitive sensor, flexible pressure sensor, hand gesture recognition, inter-day, metric learning, neural network, triplet network.

I. INTRODUCTION

H AND gesture recognition serves as an important component of human intent decoding and human-machine communication [1]. Recently, hand gesture recognition has been widely explored in sign language translation [2], augmented reality [3], and prosthetic control [4]. Previously, wearable sensors that are commonly used for gesture signal collection include: inertial measurement unit (IMU) [5], surface electromyography (sEMG) [6], pressure [7] and strain sensors [8]. With the help of artificial intelligent techniques, the gesture recognition accuracy of nearly 100% can be achieved.

However, doffing and donning of the sensors from day to day are inevitable during the daily usage. When the sensors are re-worn, displacement may cause severe performance degradation. Several studies have made efforts in alleviating the influence of inter-session and inter-day re-wearing. In [9], the authors developed an updated strategy for linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) for long-term sEMG usage, whereas in [10], a novel postprocessing algorithm which was aimed to detect and remove misclassifications of hand motions was proposed. To further improve the robustness of the gesture recognition system, large-scale sensor arrays combined with deep convolutional neural networks for data processing have been fabricated. High-density sEMG (HD-sEMG) systems, which were combined with deep learning techniques [11], had greatly alleviated the influence of electrodes shift [12], [13], and improved inter-session recognition performance [14]. In [15], the authors proposed a flexible HD-sEMG system for in-sensor adaption which received negligible affection by arm position and prolonged wearing.

Although researchers strive for solving the re-wearing performance degradation problem, most of the work has been done using sEMG and IMU. However, sEMG systems are not portable, and are susceptible to the influence of motion artifacts [16], whereas IMU sensors are more suitable for dynamic measurements, and are rigid. Both of them are not suitable for prolonged wearing and daily usage. To date, flexible sensors are emerging as new sensing methods for human-machine interface, and plenty of them are utilized in hand gesture recognition. In [17], the authors used five yarn-based triboelectric stretchable sensors which were mounted on finger joints for sign language translation. Other finger-joint-mounted
strain sensors were also explored in [18] and [19]. Except for sensing finger movements, the flexible strain sensors were also designed as a skin patch for sensing the strain distribution when gestures were performed [20]. Pressure sensors which were worn on the wrist and forearm to map the pressure distributions of the deformed tendons and muscles were also reported in [21], [22], [23], and [24]. However, inter-session and inter-day performances of these sensors have been rarely investigated.

In this paper, a flexible wristband integrated with a multi-channel capacitive array was used as a hand gesture sensing interface. The capacitive array contained $4 \times 8$ sensing pixels in total, and could be easily worn on the wrist compared to other gesture signal collection methods. When the gestures were performed, the deformation of the skin around the wrist pressed the sensing pixels inside the flexible wristband, then the pressure change was converted to capacitance change, and sampled by a customized readout circuit. Each data frame was interpolated into a high-resolution image and processed using image processing techniques. The triplet network [25], [26], which is frequently used in the context of face verification and face identification, was used as a feature extractor in the inter-day hand gesture recognition application, owing to its outstanding distinctive feature learning ability. Five days of inter-day re-wearing experiments were conducted. The performances of the triplet network and its counterpart, a convolutional neural network which was trained with softmax-cross-entropy loss, were compared. Also, except for the effect of the training loss function, the influence of the number of electrodes in the capacitive array was also investigated. The contributions of this paper are listed as follows: 1) A flexible, easy-to-wear wristband integrated with a highly sensitive capacitive array was fabricated and used as a hand gesture sensing tool. 2) The triplet network was implemented as a feature extracting technique to deal with high dimensional, image-like capacitive array data. 3) Five consecutive days of experiments of doffing and re-donning the wristband were conducted for five subjects, and exhaustive evaluations were performed.

The rest of this paper is organized as follows. Section II shows the overview of the wristband system. Section III introduces the experimental procedures, the triplet network architecture, and the evaluation methods. Results are presented in Section IV. We discuss and conclude in Section V and Section VI, respectively.

II. SENSING SYSTEM OVERVIEW

The wristband had a multi-layer structure which contained a touch bump array (with a size of $4 \times 8$), a top electrode array layer (size of $4 \times 8$), a polyethylene terephthalate (PET) spacer grid, an iontronic film with micro-structures templated from sandpaper [27], a bottom electrode array layer (with a size of $4 \times 8$), and a flexible silicone substrate. Two electrode arrays and the iontronic film in between formed a pressure sensing array with a high sensitivity and low crosstalks between channels, which can be used for detecting the pressure distribution around the wrist. The touch bump array was for concentrating pressure on each sensing pixel, and the PET spacer grid served as a supporting structure for improving the recovery ability of the sensing pixel under large initial pressure. The tightness-adjustable Velcros were fixed on the flexible silicone substrate using silicone glue. The diameter of each sensing element was 4 mm, and the spacing between each row was 7 mm. The spacing between each column varies...
from 7 mm to 14 mm as annotated in Fig. 1(a), since the
density of tendons are high at the center of the underside of
the wrist and sparse otherwise [28].

Fig. 1(a) and Fig. 1(b)-(c) show the schematic graph and the
photograph of the components of the flexible wristband, which
include: a silicone touch bump array molded from acrylic
molds, a top electrode array screen-printed on a 35-μm-thick
polyethylene terephthalate (PET) substrate, 100-μm-thick PET
films arranged into a 4 × 8 grid, polyvinyl alcohol/phosphoric
acid iontronic film casted with the height of 750 μm and
peeled off from 10000-grit sandpaper, a bottom electrode array
screen-printed on a 35-μm-thick PET substrate, and a silicone
substrate.

The capacitive array was connected to the data acquisition
system through a flexible printed circuit (FPC) connector.
The data acquisition system was composed of: eight voltage
followers for stabilizing sinusoidal signal input, eight-channel
analog switches for multiplexing, eight capacitance-voltage
converters for converting capacitance value into voltage value,
eight twelve-bit analog-to-digital converters for collecting
voltage signals, and moving average filters programmed in
the micro controller for de-noising. The filtered signals
were transmitted to a laptop through a universal synchro-
nous/asynchronous receiver/transmitter. The sampling rate
was 100 Hz.

### III. METHODS

This section describes the proposed methods in detail,
including the subjects’ information, the experimental proce-
sure, the data preprocessing method and the feature extraction
process using the triplet network. The evaluation protocol and
the statistical analysis method are also introduced.

#### A. Subjects

Five healthy and non-disabled participants were recruited
in this study. The detailed information of the subjects are
shown in Table I. All subjects were provided informed written
consents and the experiments had been approved by the Local
Ethics Committee of Peking University.

#### B. Experimental Protocol

During the experiments, the wristband was placed on the
dominant hand of the subject. The wristband was wrapped
around the wrist with the center of the capacitive array
approximately aligned with the palmaris longus, and the upper
edge of the wristband aligned with the ulnar and radius
styloid. When the gestures were performed, tendon and skin
deformation could be directly detected by the pressure sensing
array (as shown in Fig. 2).

![Fig. 2. The placement of the wristband on the wrist.](image)

Seven gestures were selected, which include: the letter
“s” (G1), one (G2), two (G3) in American sign language,
wrist flexion (G4), wrist extension (G5), ulnar deviation (G6),
and radial deviation (G7). These gestures could be linked
to meaningful control commands in remote-controlled cars,
visual-augmented reality games and so on. These gestures are
shown in Fig. 3.

In the inter-day re-wearing experiments, the data were
collected on five consecutive days. On each day, the wristband
was put on the wrist, and each gesture was performed for
5 s with a rest interval of 5 s in between. Seven gestures
were performed in a sequential order, and each run was
referred as a trial. Each trial was repeated for five times,
and this whole process was referred as a session. Then the
wristband was taken off and re-donned with experimenters
supervision, and another data collecting session began. The
data of three sessions were recorded on each experiment
day. The experimental data collection procedure is shown in
Fig. 4. It is worth noting that even though with experimenter’s
supervision, a small shift of the wristband with respect to the
original location where the data of the first day was collected
and tightness change of the wristband with respect to the wrist
were inevitable.

#### C. Data Preprocessing

The signals were sampled at 100 Hz. To remove the noise
signals, a moving average filter with a buffer size of 20 ms
was implemented.
D. Feature Extraction Networks and Classifiers

As shown in Fig. 6, the triplet network contained three identical neural networks in parallel, among which the weights were shared. The triplet network aims at learning the feature embeddings that for the samples from the same class, the extracted features were close to each other in the feature space, meanwhile for the samples from different classes, the extracted features were far apart from each other.

To achieve the distinguishing ability of feature extraction and feature separation, the triplet network took triplets of images (which included an anchor image, a positive image, and a negative image) as the input training data. The anchor image \( X^a_i \) was a randomly selected image from the data set, the positive image \( X^p_i \) was the image that had the same class label as the anchor image, and the negative image \( X^n_i \) was the image that had the different class label as the anchor image.

1) Triplet Loss: As proposed in [25], we define the mapping \( f \) between the input image and the corresponding feature embedding as follows:

\[
x_{\text{map}} = f(X)
\]  

where the feature embedding \( x_{\text{map}} \in \mathcal{R}^d \) is in a \( d \)-dimensional feature space, and is constrained to have unit \( L_2 \) norm \( \|x_{\text{map}}\|_2 = 1 \). \( f = f(\cdot; \theta) \), where \( \theta \) represents the weights of the network.

In the feature space, we want that the distance between an anchor image \( X^a_i \) of a class and all other images \( X^p_i \) from this class is less than the distance between this anchor image and any image \( X^n_i \) from any other class by a pre-defined margin:

\[
dist(f(X^a_i), f(X^p_i)) + \alpha < dist(f(X^a_i), f(X^n_i))
\]  

where \( \alpha \) is the pre-defined margin, and \( dist(\cdot, \cdot) \) is the distance metric. In this study, we took the Euclidean distance as the distance metric.

During the training process, we aimed at minimizing the triplet loss:

\[
L = \sum_{i=1}^{N} \left[ \text{dist}(f(X^a_i), f(X^p_i)) + \alpha - \text{dist}(f(X^a_i), f(X^n_i)) \right]_+
\]

where \( N \) is the cardinality of all possible triplets, and \([\cdot]_+\) denotes the ramp function.

2) Single Network Architecture: We adopted the convolutional neural network structure as the parallel identical networks in the triplet network. Each convolutional layer had a kernel size of \( 3 \times 3 \) and kernel number of \( k \), and was followed by a max pooling layer with a kernel size of \( 2 \times 2 \) and a

Fig. 4. The experimental procedure.

Fig. 5. The framework for data preprocessing. Only 80% of the data of the first day were used for training the feature extraction model, and all the rest data were used for evaluation.

To eliminate the effect of the initial pressure when the wristband was placed on the wrist, we subtracted the baseline voltage value \( V_0 \) from the current output voltage value \( V \):

\[
V_{\text{out}} = V - V_0
\]

where \( V_0 \) is defined as the average output voltage over a five-second rest period after the wristband was fixed on the wrist. For each gesture, only a 3-s clip during which the muscles and tendons stayed stable was taken as the valid data samples out of a 5-s gesture.

The thirty-two-channel signals which were filtered and calibrated were then reshaped into the shape of \( 4 \times 8 \), forming an original pressure mapping around the wrist. To estimate the pressure distribution of the areas which were not covered by electrodes, we interpolated the original pressure mapping to \( 32 \times 64 \) using bicubic interpolation method.

The final procedure of data preprocessing was normalizing each pixel in the interpolated pressure mapping using min-max normalization method:

\[
x_{\text{normalized}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}
\]

where \( x_{\text{normalized}} \) is the min-max normalized pixel value, \( x \) is the pixel value before normalization, \( x_{\text{min}} \) is the minimum value in the entire pressure mapping, and \( x_{\text{max}} \) is the maximum value in the entire pressure mapping. After normalization, each pixel in the pressure mapping had the value in the range of \( 0 - 1 \). The original and the interpolated pressure mapping corresponding to each gesture are shown in Fig. 3. All the preprocessing procedures are depicted in Fig. 5.

Fig. 6. The structure of the triplet network.
stride size of $2 \times 2$. We used rectified linear unit (ReLU) as the activation function of the output of the convolutional layer. By stacking $l$ convolutional and maxpooling layers, the entire network structure was constructed. The output of the last convolutional layer was flattened and fully connected to $n_{emb}$ neurons. The output of these $n_{emb}$ neurons was normalized to unit norm and could be perceived as the extracted embeddings by the network.

3) Classifiers: The outputs of the fully connected layer of the triplet network and the CNN were used as the extracted feature embeddings for classification. Nearest class mean classifier (NCM) [29] was selected as a baseline classifier for recognition:

$$m_c = \frac{\sum_{i \in c} x_{i}^{emb}}{N_c}$$  \hspace{1cm} (6)

where $m_c$ is the computed class mean vector of class $c$, $y_i$ is the class label of data sample $X_i$, $x_i^{emb}$ is the extracted feature vector of data sample $X_i$, and $N_c$ is the number of data samples of class $c$. The class mean of each class was computed and stored after the training process. In the test stage, a test sample was classified to the class with the smallest Euclidean distance between the extracted feature vector and the class mean vector. LDA and QDA were selected as the competitive classifiers for comparison.

4) Training Procedure: The network was trained with the batch size of 256 and a semi-hard triplet mining method [25]. Adam [30] was used as the optimization technique, and the learning rate during the entire training was set to 0.001. The network for each subject converged after 10 epochs.

The network was implemented using Python 3.8 and the Keras framework. An NVIDIA Tesla P100 GPU was used to accelerate the training process.

E. Performance Evaluation for Inter-Day Application

In order to demonstrate the performance of hand gesture recognition for inter-day application using the triplet network, two aspects were considered: the effect of the network structure and the effect of the electrode number. Details of these two aspects are elucidated in the following sections.

1) The Effect of the Network Structure: To further analyze the excellent feature extraction and the discrimination ability of triplet network, we compared the recognition result of the triplet network and its counterpart, which is the convolutional neural network (CNN) trained with softmax-cross-entropy loss.

For fair comparison, we first investigated the optimized architecture hyperparameters individually for both network structures. The number of convolutional kernel $k$, the number of convolutional layer $l$, and the output embedding feature size $n_{emb}$ were evaluated for both structures. The convolutional kernel size for both networks was set to $3 \times 3$ and stride size was set to $1 \times 1$, and each convolutional layer was followed by a maxpooling layer with a pooling size of $2 \times 2$ and a stride size of $2 \times 2$. ReLU was used as the activation function for the output of the convolutional layer. For the triplet network, $L_2$ regularization was applied to the output of the last embedding layer, and for the CNN, softmax activation was applied to the last layer and categorical cross-entropy was used as the training loss. Details of structural exploration and training parameters can be found in Table II.

The data of three sessions on the first day were used to choose hyperparameters (kernel number $k$, convolutional layer number $l$, embedding size $n_{emb}$, and classifier type). Leave-two-session-out validation was implemented for evaluating the classification result. In each run of the validation process, one session was chosen as the training set and the other two were chosen as the validation set. This process continued until all three sessions had been selected as the training set. We ran the validation process ten times to derive the average classification accuracies and standard deviations for each subject.

The tuning process proceeded as a greedy search process and started with fixing the layer number $l = 2$, the embedding size $n_{emb} = 8$, and NCM as the baseline classifier ($C = \text{NCM}$). The optimal kernel number $k^*$ was firstly searched. Then by fixing $k = k^*$, the optimal layer number $l^*$ was searched. By proceeding this procedure, we finally used $k = k^*$, $l = l^*$, $n_{emb} = n_{emb}^*$, $C = C^*$ as a set of network architecture hyperparameters.

2) Evaluation of Inter-Day Recognition Results: In the inter-day application scenario, the data collected on the first day (three sessions) were shuffled and divided randomly into training set (80% of the data) and test set (20% of the data). The training set was used for training the model, and the test set, together with the data collected on the next consecutive four days (three sessions on each day) were used to evaluate the model (as previously depicted in Fig. 3).

Due to the stochastic nature of shuffling and batching of the training data, weight initialization of the neural network, and stochastic training process on GPU, the result of a single trained model is not statistically significant [31]. Therefore, we trained the model for twenty times, removed the models with the best and worst performance, and used the mean and standard deviation to evaluate the classification performance on each test set for each subject.

3) The Effect of the Electrode Number: Another factor that contributed to the enhanced inter-day recognition performance was the number of electrodes in the capacitive array. As reported in [32], with six sensing channels, the hand gesture recognition system could only achieve the accuracy of 30.9% in the inter-day experiment. In this study, we downsized the

| Hyper-parameters | Triplet network | CNN |
|------------------|----------------|-----|
| **Adjustable**   |                |     |
| Kernel number $k$| $2,4,6,8,18$   | $2,4,6,18$ |
| Layer number $l$ | $2,3,4$        | $2,3,4$  |
| Embedding size $n_{emb}$ | $8,16,24,\ldots,96$ | $8,16,24,\ldots,96$ |
| **Fixed**        |                |     |
| Kernel size      | $3 \times 3$  | $3 \times 3$ |
| Kernel stride    | $1 \times 1$  | $1 \times 1$ |
| Pool size        | $2 \times 2$  | $2 \times 2$ |
| Pool stride      | $2 \times 2$  | $2 \times 2$ |
| Learning rate    | 0.001          | 0.001 |
| Epoch            | 10             | 10   |
| Batch size       | 256            | 128  |
| Loss function    | Triplet loss   | Cross-entropy |
output of capacitive array to investigate the effect of electrode number on the inter-day recognition accuracy.

To downsize the column resolution, we divided the $4 \times 8$ array into blocks of $1 \times 2$, $1 \times 4$, $2 \times 1$, $4 \times 1$, and $2 \times 2$, then averaged the output within each block as the final output. The final output sizes were $4 \times 4$, $4 \times 2$, $2 \times 8$, $1 \times 8$, and $2 \times 4$ respectively. The details of array division is shown in Fig. 7.

**F. Statistical Analysis**

In the comparison between the CNN and the triplet network, and in the comparison amongst the effect of different array sizes, the statistical analysis was conducted. We used the Wilcoxon signed-rank test (a non-parametric alternative to the paired t-test) and the Friedman test (a non-parametric equivalent of the repeated-measures ANOVA) [33] with a significance level of $\alpha = 0.05$.

**IV. RESULTS**

**A. Hyperparameter Tuning for the Triplet Network and the CNN**

The hyperparameter tuning results for the CNN and the triplet network are shown in Fig. 8.

As shown in Fig. 8 (a)-(d), when the triplet network was used, the optimal kernel size for all subjects ranged from 6 to 10, the optimal layer number for all subjects was 2, the optimal embedding size scattered from 24 to 88, and the optimal classifier type was the NCM classifier. The best average performance achieved in inter-session cross validation for each subject was $93.65\% \pm 3.15\%$, $89.48\% \pm 4.60\%$, $93.86\% \pm 3.07\%$, $93.35\% \pm 3.68\%$, $82.48\% \pm 7.80\%$ respectively.

Summarized from Fig. 8, the final network architectures of the CNN and the triplet network for each subject are shown in Table III.

**B. Inter-Day Recognition Results for the Triplet Network and the CNN**

In the inter-day application scenario, the hyperparameters that were selected in the previous section were used.

Fig. 9 demonstrates the classification results on the inter-day application for the CNN and the triplet network. Averaged across five days, the triplet network achieved the classification accuracy of $93.51\% \pm 8.09\%$, $91.73\% \pm 8.56\%$, $88.48\% \pm 7.51\%$, $91.45\% \pm 8.53\%$, $94.72\% \pm 6.08\%$ for each subject respectively, and the CNN achieved the classification accuracy of $87.41\% \pm 11.93\%$, $77.14\% \pm 15.40\%$, $83.01\% \pm 12.51\%$, $90.05\% \pm 8.55\%$, $85.64\% \pm 12.26\%$ for each subject respectively.

Fig. 10 shows the average confusion matrices. We plotted the average confusion matrix over five days (fifteen sessions in total) of all the subjects to visualize the classification performance on each gesture across multiple days and subjects. The triplet network could maintain the accuracies of all four wrist gestures and two finger gestures (G1 and G2) above $90\%$, and alleviate the degradation of classification accuracy of G3.

**C. The Effect of the Electrode Number on the Inter-Day Recognition Performance**

We adopted the triplet network as feature extractor in this section, since it achieved better performance in inter-day application.

Fig. 11 (a) shows the classification accuracies of all six different sizes of arrays. The average accuracy of each array size was calculated over the data of the three sessions on each day of all five subjects. When using the array with the size of $4 \times 8$, the classification accuracy was the highest in all five days. When we started downsizing the array, the accuracies in each day began to fall. The average accuracies over five days and five subjects of arrays with the size of $4 \times 8$, $4 \times 4$, $4 \times 2$, $2 \times 8$, $1 \times 8$, $2 \times 4$ were $91.98\%$, $87.51\%$, $74.45\%$, $87.58\%$, $77.24\%$, $83.92\%$, respectively.

The Friedman test of different sizes of capacitive array is summarized in a probability matrix (p matrix) as shown in Fig. 11 (b). In the matrix, p denotes the probability of accepting the null hypothesis (there does not exist significant difference between the two compared division schemes), and the results are listed in the upper right of the p matrix. There was a significant difference ($p < 0.05$) between the result of full capacitive array size and all versions of down-sampled array size.

**V. DISCUSSION**

Prolonged wearing and inter-day doffing-and-donning are inevitable for human-machine interfaces in hand gesture recognition. The question of how to avoid drastic performance degradation without re-collecting training data after days of usage and multiple times of re-wearing still remains to be
Several studies have reported inter-session and inter-day hand gesture recognition results, however, these are all based on sEMG systems, which are not convenient to wear on a daily basis. This study used a flexible multi-channel capacitive pressure sensing wristband to collect pressure distribution around wrist, and the triplet network was used as a feature extracting tool to alleviate performance degradation over days. All the evaluations were conducted in the inter-session and inter-day settings. According to the experimental results, combining the excellent feature extracting ability of the triplet network with the multi-location pressure sensing ability of the flexible wristband, the drop in classification accuracy was effectively decreased. In this section, the discussions mainly concentrate on the effects of the usage of the triplet network and the effects of electrode array size on the recognition results.
A. Comparison Between the CNN and the Triplet Network

1) Recognition Accuracies in the Inter-Session Application:
As we can observe in Fig. 8(b) and (f), increasing the number of convolutional layer rarely contributed to the increment of classification accuracy for both the CNN and the triplet network (except for the CNN trained for subject 5). This might be caused by the lack of variance in the training data and overfitting on the training sessions. From Fig. 8 (d) and (h), we can see that the baseline classifier NCM achieved the best inter-session classification result for both the CNN and the triplet network for most subjects, on the contrary, LDA and QDA, which are more complex methods than NCM, behaved worse on the recognition accuracy. This indicates that overfitting the training data could severely jeopardise the generalization ability in the inter-session re-wearing application. From the accuracy results of the final optimal architecture, we observe that the triplet network achieved higher accuracies over the CNN for most subjects (except for subject 2, the CNN was slightly better).

2) Recognition Accuracies in the Inter-Day Application: Fig. 9 and the average inter-day classification results for each subject show that, in consecutive five days of testing, the average recognition accuracies of the triplet network were higher than that of the CNN (p < 0.05 for all subjects). During the experiment, we found that the pattern of which the subject performed a gesture would greatly affect the inter-day recognition performance. Subject 2 performed G7 very differently (compared to the first day) in the first two sessions on the third day, and this resulted in the decreased recognition accuracies of 71.89% and 80.53%. Besides, malfunctioning of capacitive array channels might undermine the inter-day accuracies as well. As shown in Fig. 9, the malfunctioning of a column channel in the last session on the third day for subject 4 resulted in an accuracy of 72.97%.

In Fig. 10, we can see that the recognition accuracies on G5 and G6 using the CNN and the triplet network were roughly the same. However, the difference between the two networks was observable for the finger gestures (G1-G3). This proves the excellent embedding distribution preservation ability of the triplet network in the feature space.

The classification accuracy degradation over days was mainly caused by finger gestures. Although the re-wearing of the wristband was supervised by the experimenter, a slight electrode shift was inevitable. A slight electrode shift might not cause severe classification result declining of the wrist gestures (since the movements were large and easy to distinguish), however, it might affect the performance of finger gestures, because finger gestures involved more subtle and intricate tendon and muscle movements, thus were susceptible to minor displacement. Moreover, subjects performing gestures differently over time might also resulted in classification accuracy decreasing.

It is notable that some of the inter-day recognition accuracies of triplet network were higher than that of inter-session setting (for example, the inter-session accuracy of subject 5 was 92.66%, whereas the inter-day recognition accuracy of three sessions on the second day were 99.41%, 97.35%, and 96.41% respectively). This might be caused by the enriched training data for model training in the inter-day application (all three sessions on the first day were used for training).

3) Feature Distributions Using the CNN and the Triplet Network in the Inter-Day Application: To evaluate the feature distributions of feature embeddings extracted by the CNN and triplet network after five days, we selected the representative examples of G1 in the second session on the last experimental day of subject 1. The recognition accuracies of G1 when model was trained with the CNN and the triplet network were 41.07% and 100.0%, respectively. We can also observe that for the CNN, samples from G1 were mainly misclassified to G6, with the percentage of 39.88%. To find out the reasons that caused this misclassification for the CNN, we first visualized the distance between positive pairs and the distance between negative pairs for both CNN and triplet network (Fig. 12 (a)-(b)).

Five-thousand positive pairs were randomly selected from all the possible combinations of samples in G1, and five-thousand negative pairs were randomly selected from all the possible combinations of sample pairs in G1 and G6 (one sample from G1 and the other from G6). All the selected samples were pre-processed and passed through the trained feature extractor (CNN or triplet network). The distances between each positive pairs and between each negative pairs were computed, and the density distributions were estimated. From Fig. 12 (a)-(b), we can see that the separation ability between G1 and G6 of the triplet network was better than that of the CNN. However, according to the density distribution, features of G1 and G6 were separable for both CNN and triplet network. Therefore, the misclassification between G1 and G6 using the CNN was not caused by the confusion between positive and negative pairs.
To further investigate this phenomenon, we also visualized the density distributions of distance (denoted as $d_p$) between features of G1 and centroid of G1 (computed by the NCM classifier), and the density distributions of distance (denoted as $d_a$) between features of G1 and centroid of G6. As illustrated in Fig. 12 (c)-(d), for features extracted by the CNN, the density distributions of $d_p$ and $d_a$ are severely overlapped, which means that G1 and G6 were not separable by using the NCM classifier. Whereas for features extracted by the triplet network, the density distributions of $d_p$ and $d_a$ could be clearly separated. This means that the triplet network can retain the geometric distributions of features in the feature space.

4) Computational Time: There were approximately 25200 training samples in total for the inter-day application for each subject. The average training time for the CNN (10 epochs) and the triplet network (10 epochs) were 10.05 s and 24.85 s individually. All the training were run on NVIDIA Tesla P100 GPU. To measure the testing time for each data sample, we input the testing data of one session and then compute the average testing time for each sample. When the tests were run on an Intel Core i7 CPU, the average testing time of each sample using the CNN and the triplet network were 0.35 ms and 0.44 ms respectively.

Although the training time of the triplet network took two times longer than that of the CNN, the training procedure only took place on the first day, and 25 s is acceptable for data training. When running the testing process, the CNN and the triplet network achieved similar recognition time of less than 0.5 ms, which is sufficient for real-time application.

B. The Effect of the Number of Electrodes

Although hand gesture recognition using pressure sensors had been reported in [21], [22], [23], and [24], pressure sensors with large array size are rarely used in gesture recognition tasks. Most of the studies reported wrist-worn and forearm-worn armband with less than ten channels [21], [22], [23], [32], which were integrated with pressure sensors with low sensitivity. In [24], the authors demonstrated a flexible capacitive wristband with fifteen sensing channels, however, this was still not accurate enough to map the pressure distribution around the wrist. In this study, we found that the number of sensing electrodes also contributed to improving the classification accuracy in the inter-day application.

As shown in Fig. 11 (a), using $4 \times 8$ capacitive array yielded the best classification results in all five test days. The classification performance when using full size array ($4 \times 8$) can also be confirmed in Fig. 11 (b), from where we can observe that, in a pair-wise comparison scheme, the capacitive array with the size of $4 \times 8$, had a significant difference ($p < 0.05$) between other down-sampled array. This means that although down-sampling the array size can simplify the wristband fabrication process and readout circuit design, the inter-day classification results have to be compromised.

Some other interesting facts can also be found in Fig. 11 (b). If we down-sampled the array size from thirty-two channels to sixteen channels ($2 \times 8$ and $4 \times 4$), the average performance decreased about 4%, and these two division schemes showed no significant differences ($p = 0.636$). However, if we further down-sampled the array size to eight channels ($4 \times 2$, $1 \times 8$, and $2 \times 4$), the average performance sharply declined, and significant differences existed amongst these division schemes. This observation also proves the importance of the array size.

C. Limitations and Future Works

Although in this paper, the purpose of inter-day hand gesture recognition has been fulfilled using a flexible capacitive wristband and triplet network, there are still some limitations.

Firstly, although inter-day performance of the flexible wristband has been enhanced by using multi-channel sensing array and triplet network, the classification accuracies over days were not satisfactory, especially compared to hand gesture recognition systems using sEMG. Therefore the wristband system has to be further improved. For example, fabricating sensor arrays with high density and low inter-channel crosstalks could alleviate the performance drop caused by electrode shift, and integrating adaptive and continual learning algorithms into the system architecture could mitigate the influence of sensor performance degradation and gesture pattern change over days.

Secondly, the data were analyzed in an offline manner, and the system has not been connected to the controlling module. Therefore in the future, we will implement the algorithm online for real-time hand gesture recognition, and the recognized gestures will be used as controlling commands, which will benefit the robust and prolonged wearing of gesture recognition interfaces in the applications of remote operation and virtual reality.

VI. CONCLUSION

In this paper, we proposed and validated the feasibility of a highly sensitive multi-channel capacitive array wristband in the inter-day hand gesture recognition application using...
a triplet network. The pixelated outputs of the capacitive array were considered as frames of images, and mapped the pressure distributions around the wrist into patterns in the image. Being enhanced by the triplet network for feature embedding, the inter-day performance of the wristband for gesture recognition was further improved. From the experimental results, we can draw the following conclusions. First, the triplet network, which contained three identical CNN structures in parallel and adopted triplet loss during training, outperformed the conventional CNN by 7.33% (p < 0.05) in average classification accuracy in the inter-day re-wearing context. Second, the capacitive array size highly influenced the recognition results. The full-size array (4 x 8) yielded much better results than a down-sized array (p < 0.05). This study provides a new methodology that directly measures the pressure mapping around the wrist through an easy-to-wear, triplet network-enhanced flexible capacitive wristband, which provides us with new insights to the solutions of inter-day re-wearing of human-machine interfaces for hand gesture recognition.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for the valuable suggestions on improving the manuscript. They would like to thank Shichang Zhang for the circuit fabrication and Yuwen Lu for the advice on the wristband design. They also like to thank Xinzhao Guo, Yuwen Lu, Zilu Wang, Zhaoyang Wang, and Qi Tang for their kind assistance in completing the experiments.

REFERENCES

[1] M. J. Cheok, Z. Omar, and M. H. Jaward, “A review of hand gesture and sign language recognition techniques,” Int. J. Mach. Learn. Cybern., vol. 10, no. 1, pp. 131–153, Jan. 2019.

[2] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang, and L. Yuan, “Hand gesture recognition using deep feature fusion network based on wearable sensors,” IEEE Sensors J., vol. 21, no. 1, pp. 539–547, Jan. 2021.

[3] F. Wen et al., “Machine learning glove using self-powered conductive superhydrophobic triboelectric textile for gesture recognition in VR/AR applications,” Adv. Sci., vol. 7, no. 14, Jul. 2020, Art. no. 2000261.

[4] G. Gu et al., “A soft neuroprosthetic hand providing simultaneous myoelectric control and tactile feedback,” Nature Biomed. Eng., pp. 1–10, 2021, doi: 10.1038/s41551-021-00767-0.

[5] M. Kim, J. Cho, S. Lee, and Y. Jung, “IMU sensor-based hand gesture recognition for human-machine interfaces,” Sensors, vol. 19, no. 18, p. 3827, Sep. 2019.

[6] P. Tsinganos, B. Cornelis, J. Cornelis, B. Jansen, and A. Skodras, “Improved gesture recognition based on sEMG signals and TCN,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 1169–1173.

[7] P. G. Jung, G. Lim, S. Kim, and K. Kong, “A wearable gesture recognition device for detecting muscular activities based on air-pressure sensors,” IEEE Trans. Ind. Informat., vol. 11, no. 2, pp. 485–494, Apr. 2015.

[8] S. Sun et al., “Gesture recognition based on nano-gold flexible sensor using different training modes,” in Proc. IEEE Int. Conf. Cyborg Bionic Syst. (CBS), Sep. 2019, pp. 72–77.

[9] X. Chen, D. Zhang, and X. Zhu, “Application of a self-enhancing classification method to electromyography pattern recognition for multifunctional prosthesis control,” J. Neuroeng. Rehabil., vol. 10, no. 1, pp. 1–13, Dec. 2013.

[10] S. Amsuss, P. M. Goebel, N. Jiang, B. Graimann, L. Paredes, and D. Farina, “Self-correcting pattern recognition system of surface EMG signals for upper limb prosthesis control,” IEEE Trans. Biomed. Eng., vol. 61, no. 4, pp. 1167–1176, Apr. 2013.

[11] W. Geng et al., “Gesture recognition by instantaneous surface EMG images,” Sci. Rep., vol. 6, no. 1, pp. 1–8, 2016.

[12] L. Wu, X. Zhang, K. Wang, X. Chen, and X. Chen, “Improved high-density myoelectric pattern recognition control against electrode shift using data augmentation and dilated convolutional neural network,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 28, no. 12, pp. 2637–2646, Dec. 2020.

[13] R. Hu, X. Chen, X. Zhang, and X. Chen, “Adaptive electrode calibration method based on muscle core activation regions and its application in myoelectric pattern recognition,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 29, pp. 11–20, 2020.

[14] Y. Du, W. Jin, W. Wei, Y. Hu, and W. Geng, “Surface EMG-based inter-session gesture recognition enhanced by deep domain adaptation,” Sensors, vol. 17, no. 3, p. 458, 2017.

[15] A. Moin et al., “A wearable biosensing system with in-sensor adaptive machine learning for hand gesture recognition,” Nature Electron., vol. 4, pp. 54–63, Jan. 2021.

[16] K. Lienhard et al., “SEMG during whole-body vibration contains motion artifacts and reflex activity,” J. Sports Sci. Med., vol. 14, no. 1, p. 54, 2015.

[17] Z. Zhou et al., “Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays,” Nature Electron., vol. 3, no. 9, pp. 571–578, Sep. 2020.

[18] G. Gu et al., “Integrated soft ionotronic skin with stretchable and transparent hydrogel-elastomer ionic sensors for hand-motion monitoring,” Soft Robot., vol. 6, no. 3, pp. 368–376, 2019.

[19] M. Wang et al., “Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors,” Nature Electron., vol. 3, pp. 563–570, Sep. 2020.

[20] S. Jiang, L. Li, H. Xu, J. Xu, G. Gu, and P. B. Shull, “Stretchable e-skin patch for gesture recognition on the back of the hand,” IEEE Trans. Ind. Electron., vol. 67, no. 1, pp. 647–657, Jan. 2019.

[21] X. Liang et al., “Fusion of wearable and contactless sensors for intelligent gesture recognition,” Adv. Intell. Syst., vol. 1, no. 7, Nov. 2019, Art. no. 1900088.

[22] Y. Zhang, B. Liu, and Z. Liu, “Recognizing hand gestures with pressure-sensor-based motion sensing,” IEEE Trans. Biomed. Circuits Syst., vol. 13, no. 6, pp. 1425–1436, Dec. 2019.

[23] X. Liang, R. Ghannam, and H. Heidari, “Wrist-worn gesture sensing with wearable intelligence,” IEEE Sensors J., vol. 19, no. 3, pp. 1082–1090, Feb. 2018.

[24] H. Truong et al., “CapBand: Battery-free successive capacitance sensing wristband for hand gesture recognition,” in Proc. 16th ACM Conf. Embedded Networked Sensor Syst., Nov. 2018, pp. 54–67.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 815–823.

[26] E. Hoffer and N. Ailon, “Deep metric learning using triplet network,” in Proc. Int. Workshop Similarity-Based Pattern Recognit., 2015, pp. 84–92.

[27] N. Bai et al., “Graded inaffilable architecture-based iontronic pressure sensor with ultra-broad-range high sensitivity,” Nature Commun., vol. 11, no. 1, pp. 1–9, Dec. 2020.

[28] P. B. Shull, S. Jiang, Y. Zha, and X. Zhu, “Hand gesture recognition and finger angle estimation via wrist-worn modified barometric pressure sensing,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 4, pp. 724–732, Apr. 2019.

[29] A. R. Webb, Statistical Pattern Recognition. Hoboken, NJ, USA: Wiley, 2002.

[30] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[31] D. Zhuang et al., “Randomness in neural network training: Characterizing the impact of tooling,” in Proc. Mach. Learn. Syst., 2022, pp. 316–336.

[32] E. Zheng, J. Mai, Y. Liu, and Q. Wang, “Forearm motion recognition with noncontact capacitive sensing,” Frontiers Neurorobotics, vol. 12, p. 47, Jul. 2018.

[33] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” J. Mach. Learn. Res., vol. 7, pp. 1–30, Dec. 2006.