Robotica (2023), **41**, pp. 775–788
doi:10.1017/S026357472200159X

RESEARCH ARTICLE

Time–frequency feature transform suite for deep learning-based gesture recognition using sEMG signals

Xin Zhou[1,2,†], Jiancong Ye[1,3,†], Can Wang[1,*], Junpei Zhong[4] and Xinyu Wu[1]

[1]Guangdong Provincial Key Lab of Robotics and Intelligent System, Shenzhen Institute of Advanced Technology, Chinese Academy of Science, Shenzhen 518055, China, [2]University of Science and Technology of China, Hefei, Anhui 230026, China, [3]Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou 511442, China, and [4]The Hong Kong Polytechnic University, Hong Kong, China

[*]Corresponding author. E-mail: can.wang@siat.ac.cn

Received: 8 June 2022; **Revised:** 13 September 2022; **Accepted:** 6 October 2022; **First published online:** 4 November 2022

Keywords: sEMG, gesture recognition, deep learning, neural networks, TFFT suite

Abstract

Recently, deep learning methods have achieved considerable performance in gesture recognition using surface electromyography signals. However, improving the recognition accuracy in multi-subject gesture recognition remains a challenging problem. In this study, we aimed to improve recognition performance by adding subject-specific prior knowledge to provide guidance for multi-subject gesture recognition. We proposed a time–frequency feature transform suite (TFFT) that takes the maps generated by continuous wavelet transform (CWT) as input. The TFFT can be connected to a neural network to obtain an end-to-end architecture. Thus, we integrated the suite into traditional neural networks, such as convolutional neural networks and long short-term memory, to adjust the intermediate features. The results of comparative experiments showed that the deep learning models with the TFFT suite based on CWT improved the recognition performance of the original architectures without the TFFT suite in gesture recognition tasks. Our proposed TFFT suite has promising applications in multi-subject gesture recognition and prosthetic control.

1. Introduction

Surface electromyography (sEMG) is a technique involving the use of electrodes overlying a muscle to capture biologic signals from electrical muscle activity. Recently, sEMG signals have been widely used in several applications in electrophysiological studies, including movement intention recognition [1] and angle prediction [2–4]. This wide usage results from the sufficient latent biological information provided by sEMG. Meanwhile, several sEMG databases such as Ninapro [5] and MyoUp [6] have been made publicly available as benchmarking tools for studying the relationships between sEMG and kinematics. The popularity of public benchmark datasets enables the easy validation and comparison of proposed methods with other existing methods, which is helpful for facilitating the progress of relevant research.

Existing methods based on sEMG data for gesture classification tasks in electrophysiological studies can be characterised in several basic steps: (1) pre-processing sEMG data to minimise noise; (2) extracting hand-crafted features such as time-domain and frequency-domain features, or capturing feature representations automatically by creating deep learning (DL) models; and (3) completing gesture classification tasks using extracted features. Many state-of-the-art models based on DL have achieved superior performance in gesture recognition [7–10]. However, owing to differences in the muscle strengths, fat contents, and skin impedances of different subjects, sEMG signals almost vary with persons even on the same motion. Some studies have focused on training a special model for each participant [11], which is computationally costly.

[†]These authors contributed equally to this work and should be considered co-first authors.

https://doi.org/10.1017/S026357472200159X Published online by Cambridge University Press

Adding artificial prior knowledge to a neural network can often provide guidance for tasks [12, 13]. However, the types of prior knowledge suitable for specific tasks require clarification. For example, we believe that sEMG signals obtained from a human in a resting state can provide sufficient information for characterising myoelectric properties as a prior for describing features of sampled muscles, which is crucial for eliminating the discrepancy in the biological data of different individuals. Specifically, we attempted to use DL models to capture features of the sEMG signals under different motion modes and participants. Meanwhile, an external DL model is used for obtaining relevant coefficients to fine-tune the original features.

In this study, we designed a neural network-based suite named **time–frequency** features transform (TFFT). As no previous work has investigated how to obtain feature priors from different participants for incorporation into signal feature extraction, we explored the possibility of using the time–frequency diagrams of continuous wavelet transform (CWT) as the priors. We believe that these diagrams can encapsulate rich prior, as our experiments show. Specifically, such time–frequency diagrams are converted from multiple groups of sEMG signals produced by individuals in a resting state and packaged into an image matrix with corresponding channels. The proposed suite is then conditioned on the multichannel time–frequency diagram to generate a pair of modulation parameters so that the features of the network can be affine transformed. In summary, our contributions are as follows:

1. We proposed a DL-based time–frequency feature transform suite for gesture recognition that uses CWT to extract subjects' prior knowledge to improve recognition accuracy in multi-subject scenarios.
2. The effectiveness of the TFFT suite was validated on the open dataset Ninapro DB8. Specifically, TFFT was equipped on LSTM and CNN and compared with previous state-of-the-art models and classical methods using temporal features. Experimental results show that LSTM and CNN with TFFT achieve higher accuracy and validate the effectiveness of our proposed method.

2. Related work

2.1. Surface electromyography and deep learning methods

Many studies have introduced DL methods for intention recognition or angle estimation based on sEMG signals. Early methods explored the application of the convolutional neural network (CNN) architecture in gesture recognition using sEMG [14], which is the first DL-based architecture applied to sEMG signals to classify data from the Ninapro database for improved performance compared with that of support vector machine. Another modified CNN architecture called LeNet was used to classify 50 hand movements on the Ninapro database [15]. The classification accuracy and robustness of the CNN structure were better than those of various machine learning techniques, including linear discriminant analysis, support vector machine, and k-nearest neighbour. These methods showed that CNNs have desirable representation ability and perform well in sEMG-based movement recognition tasks. Moreover, many studies have focused on the recurrent neural network (RNN), which is superior in processing temporal series. Additionally, several types of RNN-based DL algorithms, such as long short-term memory (LSTM) and gated recurrent unit, have been used to extract temporal information from sEMG [16, 17] while DL has been combined with certain schemes such as attention mechanism and machine learning models for improved performance [18–20].

Contemporary sEMG algorithms are increasingly becoming more DL-based methods for learning the mapping of sEMG signals, and they output results in an end-to-end manner. Only a few studies have introduced prior information to make DL more robust for addressing this problem. In this study, we explore feature priors in the form of time–frequency diagrams in a neural network framework. The feature extraction and transformation parameter learning of the time–frequency map are also based on time and frequency scales.

https://doi.org/10.1017/S026357472200159X Published online by Cambridge University Press

2.2. Multiuser solutions based on sEMG signals

Several studies have attempted to address the limitations of existing models in maintaining acceptable performance on sEMG signals from different individuals. In addition, the heterogeneity of sEMG data from different individuals hinders performance improvement. Here, we review related solutions that are based on sEMG signals.

Matsubara *et al*. [21] proposed a bilinear model that takes a user-dependent factor as one of two linear factors. They used an adaptive method to estimate the user-dependent factor, enabling the bilinear sEMG model to extract user-independent features from new user data. The extracted features become less relevant to users. Xiong [22] hypothesised that the non-stationary and complex waveform of sEMG signal can be decomposed to a limited number of motor unit action potentials (MUAPs) with distinct weight values. The authors created a model to identify the MUAPs for different individuals. Several studies have introduced canonical correlation analysis (CCA) technology for extracting the latent correlations between different sets. Khushaba *et al*. [23] proposed a framework for multiuser sEMG interfaces using CCA to address the data heterogeneity of different users. Xue *et al*. [24] proposed a framework based on CCA and optimal transport, further reducing the discrepancies in data distribution between the transformed training and test sets.

These studies highlighted above focused on eliminating individual differences of sEMG signals from many manual modes including manual feature selection, model matching, and feature transformation. Our work differs from this work in two main aspects. First, we propose an effective DL component to reduce the data heterogeneity in a single forward pass conditioned on the prior with meaningful information. Our proposed TFFT network can generate feature transformed matrix and perform feature-wise manipulation adaptively. Finally, we combine LSTM and CNN with TFFT to achieve state-of-the-art gesture recognition performance.

3. Methodology

3.1. Overview

The generic structure of our proposed TFFT suite is shown in Fig. 1. It is necessary to generate a subject-specific prior map from a CWT, which is then processed by a branch network to obtain a pair of modulation parameter pairs. The modulation parameters generated by the branch network are then added to each layer of the backbone network by the affine transform. Finally, the predictions are obtained from these backbone networks.

3.2. Continuous wavelet transform

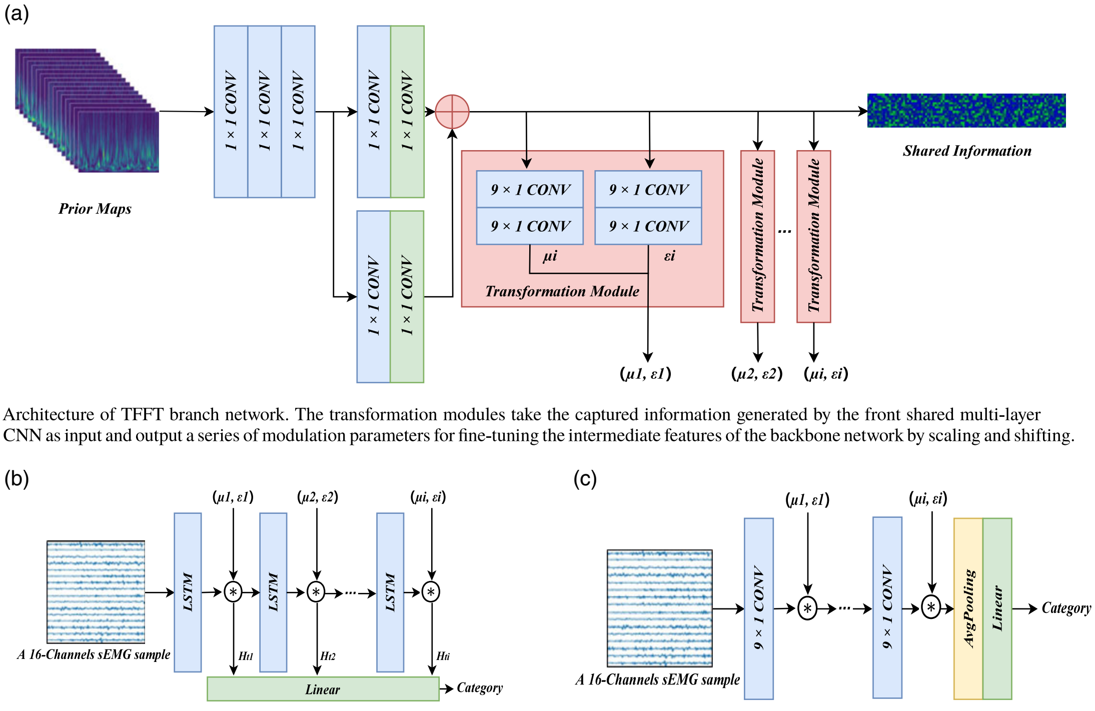
To generate prior maps, we used the CWT method for the sEMG signals of the related muscles at the resting state. The CWT of signal $x(t)$ is defined as [25]:

$$CWT_x(a,b) = |a|^{-\frac{1}{2}} \int_0^T x(t)\psi^*\left(\frac{t-b}{a}\right) dt, \tag{1}$$

$$\int_{-\infty}^{\infty} \psi(t)dt = 0, \tag{2}$$

where $\psi(t)$ represents the mother wavelet while a and b refer to scale and translation parameters, respectively. As wavelet transformation can describe the local property of a signal, the generated CWT maps can be the prior that provides necessary information about the subjects.

Specifically, we extract a series of segments of sEMG signals of each participant in the resting state, which is performed by adopting ‘Cgau8’ wavelet [26]. The CWT method is then applied to each frame. Figure 2 illustrates the processing of the sEMG signals of one participant whose muscles are in the resting state using the CWT operation to generate prior maps.



Architecture of TFFT branch network. The transformation modules take the captured information generated by the front shared multi-layer CNN as input and output a series of modulation parameters for fine-tuning the intermediate features of the backbone network by scaling and shifting.

Architecture of LSTMs with modulation parameter pairs.

Architecture of CNNs with modulation parameter pairs.

Figure 1. *Time–Frequency feature transform networks.*

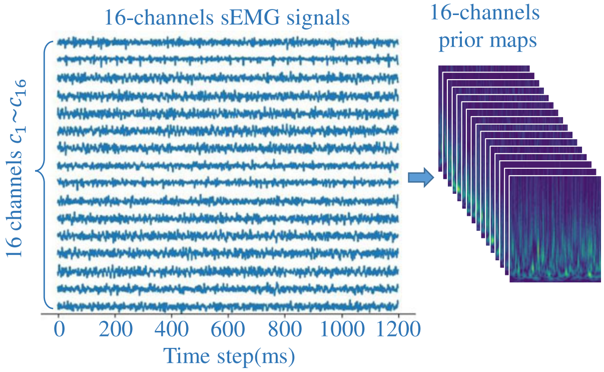


Figure 2. *The 16-channel sEMG signals that are operated by ”Cgau8” wavelet and CWT to generate a 16-channel prior maps.*

3.3. Time–frequency feature transform

The time–frequency feature transform network takes the prior maps Γ obtained by CWT as input to learn a mapping function that outputs a modulation parameter pair (μ, ε). The parameter pair adaptively influences the outputs through an affine transformation on each intermediate feature map in the neural network. The mapping function $\mathcal{F} : \Gamma \Rightarrow (\mu, \varepsilon)$ can be divided into two parts:

1. A mapping function $\mathcal{F}_\alpha : \Gamma \Rightarrow \Theta$ is used to extract shared information from the input prior.
2. Each transformation module has a unique pair of modulation parameters (μ_i, ε_i) determined by the mapping function $\mathcal{F}_\beta : \Theta \Rightarrow (\mu, \varepsilon)$.

https://doi.org/10.1017/S026357472200159X Published online by Cambridge University Press

Consequently,

$$(\mu, \varepsilon) = \mathcal{F}_\beta(\Theta) = \mathcal{F}_\beta(\mathcal{F}_\alpha(\Gamma)) = \mathcal{F}(\Gamma). \tag{3}$$

Subsequently, the parameter pair (μ, ε) is used to transform the feature map $\boldsymbol{F}$ by scaling and shifting, which is expressed as:

$$\boldsymbol{F} \circledast (\mu, \varepsilon) = (1 + \mu) \odot \boldsymbol{F} + \varepsilon, \tag{4}$$

where $\boldsymbol{F}$ has the same dimension as that of μ and ε, $\circledast$ refers to the scaling and shifting operation, and $\odot$refers to the element-wise multiplication. Therefore, a transformation module performs feature-wise manipulation according to the external prior and branch network. The architecture of the TFFT branch network is illustrated in Fig. 1(a). The normalisation operations and dropout layers behind the CNN layers are simplified. Here, we focus on the introduction of the external conditioning part. We use a dual-branch CNN for $\mathcal{F}_\alpha$ to capture necessary information from priors so that it can also be trained together with the backbone network in an end-to-end manner. In addition, the branch network can generate shared intermediate conditions Θ that can be broadcasted to all transformation modules for efficiency. Each transformation module contains separate small convolution layers that further adapt the shared Θ to gain specific parameters μ and ε.

3.4. Original architecture with TFFT

To normalise the performance of the proposed TFFT suite, the neural network framework consists of two streams: a prior branch network and a backbone network. The TFFT branch network takes multi-channel time–frequency diagrams of sEMG signals generated by CWT in the resting state as input, which are then processed by five convolutional layers to capture useful features. The extracted prior features are then shared by all transformation modules. We use three types of kernels in the prior branch network. The 1 × 1 kernel restricts the receptive field of the convolutional network for the three forward convolutional layers. However, the $1 \times H$ and $W \times 1$ kernels are used for integrating information in the scales of time and frequency.

We select the basic LSTMs or CNNs architecture as backbone networks, in which the intermediate features take scaling and shifting operations according to the modulation parameter pair μ and ε. We apply the *Softmax* function and set a fully connected layer as the classifier to recognise gesture. Although we only attempt two common architectures for the backbone network, we believe their variants are applicable and can be combined with the TFFT suite to improve performance.

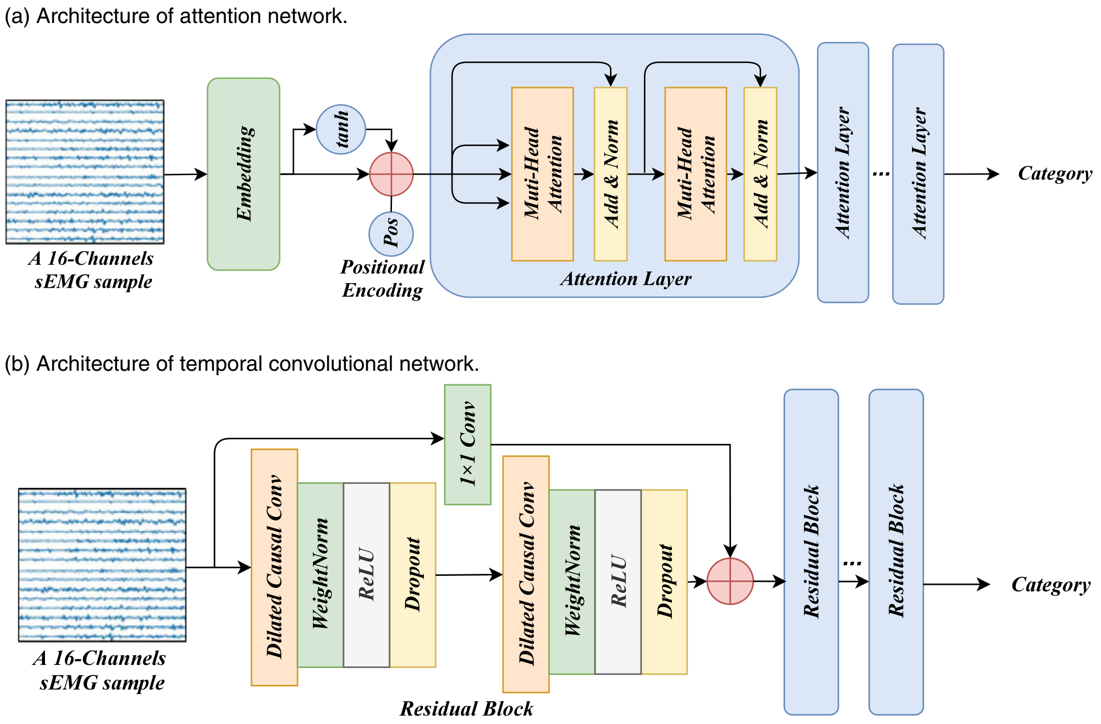
If we take the LSTM architecture as the backbone network, the output $\boldsymbol{F}_i$ of each LSTM layer would perform scaling and shifting with the parameter pair (μ_i, ε_i). Thus, we would concentrate on the output hi of the last time step of each layer, followed by a fully connected layer as a classifier to infer category. Otherwise, if we apply the CNNs architecture as the backbone network, several convolution layers would be considered, and the output would perform the operation of (4), as well as a fully connected layer as the classifier. As illustrated in Fig. 1(b) and (c), we omit all normalisation operations and dropout layers behind the LSTM and CNN layers to show the architecture of backbone networks clearly.

4. Other state-of-the-art frameworks

We compared the TFFT suite with two state-of-the-art structures for gesture recognition: DL models equipped with attention mechanisms [27] and temporal convolutional network (TCN) [28]. The detailed structure is illustrated in Fig. 3(a) and (b).

The strong modelling ability of the attention mechanism in time series data results from assigning different weight coefficients to inputs at different positions. The weight increases with increasing degree of correlation. First, the input data $\boldsymbol{X}$ go through the embedding layer to expand their feature dimension for consistency with the hidden dimension. After the embedding, an activation function, a residual connection, and positional encoding are required. The embedding layer, activation function, and position

https://doi.org/10.1017/S026357472200159X Published online by Cambridge University Press


(a) Architecture of attention network.

(b) Architecture of temporal convolutional network.

Figure 3. *State-of-the-art networks.*

encoding use the fully connected layer, *tanh*, sine and cosine functions, respectively.

$$\boldsymbol{E} = Embedding(\boldsymbol{X}), \tag{5}$$

$$\boldsymbol{Y} = Activation(\boldsymbol{E}) + \boldsymbol{E}, \tag{6}$$

$$\boldsymbol{M} = PositionalEncoding(\boldsymbol{Y}), \tag{7}$$

$$\boldsymbol{Q} = \boldsymbol{M}\boldsymbol{W}^{Q}, \boldsymbol{K} = \boldsymbol{M}\boldsymbol{W}^{K}, \boldsymbol{V} = \boldsymbol{M}\boldsymbol{W}^{V}, \tag{8}$$

where $\boldsymbol{W}^{*} \in \mathbb{R}^{d \times d}$ is weight matrix, and d is hidden dimension.

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{d}}\right)\boldsymbol{V}, \tag{9}$$

$$MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_{\mathrm{h}})\boldsymbol{W}, \tag{10}$$

where $\mathbf{h}_i = \mathrm{Attention}(\boldsymbol{Q}\mathbf{W}_{\mathrm{i}}^{\mathrm{q}}, \boldsymbol{K}\mathbf{W}_{\mathrm{i}}^{\mathrm{k}}, \boldsymbol{V}\mathbf{W}_{\mathrm{i}}^{\mathrm{v}})$, $\boldsymbol{W}$, $\boldsymbol{W}_{\mathrm{i}}^{*}$ are parameter matrices. After Eqs. (5)–(7), the three input matrices of the self-attention layer are obtained. In Eqs. (9) and (10), $\boldsymbol{Q}\boldsymbol{K}^{T}$ computes the attention score, which is divided by $\sqrt{d}$ to scale. The multi-head attention mechanism is used to parallelise the computation and focus on information at different locations.

TCN also has a good performance in time series modelling; information obtained by the current node only includes the previous node information, which is logically reasonable. However, a new problem would arise. As the length of the input sample data increases, covering the entire receptive field would be at the expense of increasing the number of network layers. Consequently, the number of network layers would maintain a linear relationship with the length of the input sample. Dilated convolution is introduced to alleviate this problem. Without sacrificing the receptive field, the number of network layers

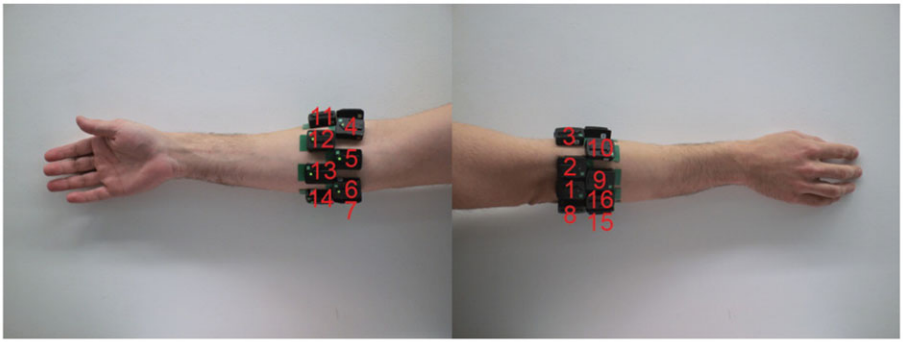


Figure 4. *Sensor layout of NinaPro DB8.*

maintains a logarithmic relationship with the length of the input data. Additionally, the introduction of weight normalisation not only increases the computational efficiency, it also counteracts the gradient explosion. Meanwhile, the activation function increases the nonlinear fitting ability of the network while the dropout operation improves the generalisation ability.

Both structures are stacked with four layers, consistent with the TFFT framework. In the training of our model, we apply the cross-entropy loss, which is widely used in the classification problem. In multiple classification problems, the cross-entropy loss is computed by the following equation:

$$CELoss(p|q) = \frac{1}{M}\sum_{1}^{M}\sum_{i=1}^{C} p_i log(q_i), \tag{11}$$

where C represents the number of categories; M represents the number of samples; and p and q represent the real and prediction probabilities, respectively.

5. Experiments and results

5.1. Experimental setup

5.1.1. Datasets

For the experiments, we used the NinaPro DB8 database, which comprises nine movements, including single-finger and functional movements: thumb flexion/extension, thumb abduction/adduction, index finger flexion/extension, middle finger flexion/extension, combined ring and little fingers flexion/extension, index pointer, cylindrical grip, lateral grip, and tripod grip. The database also contains the sampling data of the finger at rest. The creators of the database applied 16 active double-differential wireless sensors to record sEMG signals in correspondence to the radiohumeral joint of the right hand (see Fig. 4) without targeting specific muscles. The sEMG signals were sampled at a rate of 1,111 Hz and denoised in advance by data providers. Ten physically healthy and two right-handed transradial amputee participants were enlisted to obtain the dataset. The participants were asked to repeat nine movements, each movement lasting for 6–9 s and consecutive trials interrupted by 3 s of rest.

5.1.2. Sliding window

Open databases are mostly pre-processed, as reported in relevant papers and data source websites. To fully exploit open data and obtain acceptable results, we use the sliding window method to generate time slice data as a sample. Given an original sEMG sequence, $\boldsymbol{X} = [x_1, x_2, \ldots, x_t]$, which is resampled at a sampling rate $\upsilon = 500$ Hz, and the sliding window method aims to segment $\boldsymbol{X}$ with a length of $\boldsymbol{L}$ milliseconds (window size $\boldsymbol{W} = \frac{L}{2}$). The step size of the sliding windows is set to 16 ms. Figure 5 presents the segmentation and combination of multi-channel sEMG signals. The converted sEMG samples are $\boldsymbol{R}^{\boldsymbol{C}} \times \boldsymbol{W}$, where $\boldsymbol{C}$ refers to the number of electrodes.

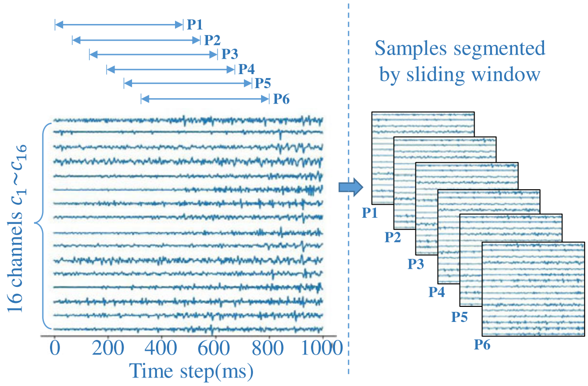

P1
P2
P3
P4
P5
P6
Samples segmented by sliding window
16 channels $c_1 \sim c_{16}$
0
200
400
600
800
1000
Time step(ms)
P1
P2
P3
P4
P5
P6

Figure 5. *Segmentation of the sEMG signals to a series of samples by the sliding window method. p_i represents a segment of sEMG signals at time i; c_e represents sEMG signals sampled by electrodes e.*

The length of the window represents the time latency of prediction with past status. To test the performance of the proposed algorithm under different time latencies in this study, we varied the window size. The step size of the sliding window is also fixed at 16 ms.

5.1.3. Preparation

After acquiring the NinaPro DB8 [29] dataset, we prepared the data for our validation experiments. As we focused on the gesture recognition of healthy persons, we only downloaded data of 10 healthy individuals and removed the data that were corrupted during the downloading. Finally, the data of eight individuals were retained; the data of each individual contained 16 channels. As illustrated in Fig. 6, the preparation process was divided into three steps:

1. We segmented the data according to their labels. The data segment with a label of zero was separated from that with other labels (the zero label indicates a resting state).
2. The data were resampled at a rate of 500 Hz using the sliding window method with a certain window length. The window length affects the recognition performance of the system. Therefore, we used window lengths of 200 ms, 300, 400, and 500 ms to segment the data. The data of all categories were then split into training and test sets in ratio 7:3.
3. Finally, the data labelled zero were resampled at a rate of 500 Hz and extracted under each subject through a window with a length of 2,400 ms. The sEMG signal segment of 16 channels was transformed by CWT to generate 16-channel prior maps.

5.1.4. Training setting

In the model training state, the number of training iterations was set to 120. We used the Adam optimiser [30], with $\beta_1 = 0.9$. The learning rate was set to 1×10^{-3} and then decayed by a factor of 10 in 60 and 100 epochs. Other hyperparameter settings are presented in Table I.

5.2. Qualitative evaluation

5.2.1. Comparison with original architecture approaches

We compared the proposed hybrid architectures based on the TFFT suite with approaches based on the original architecture, such as CNNs and LSTMs. Our proposed TFFT-based models are TFFT-CNNs and TFFT-LSTMs. To evaluate the advantages of the proposed TFFT suite, we ensured that all original architectures had four-layer depths. These four models take the sEMG signals that are self-processed

https://doi.org/10.1017/S026357472200159X Published online by Cambridge University Press

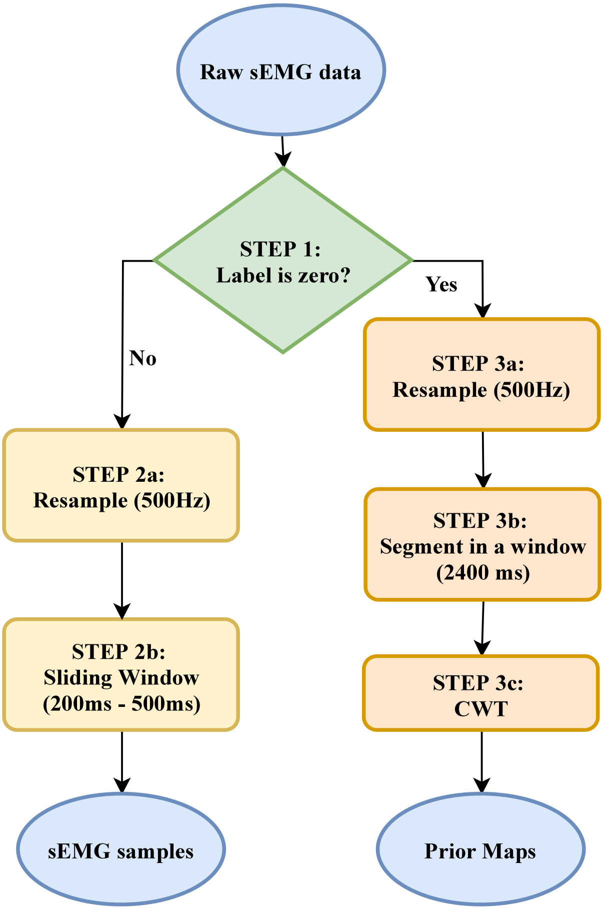


Figure 6. *Flowchart of data preparation process after acquiring NinaPro DB8 dataset.*

according to the operations described in Section 5.1.3 as input, which are then trained on the training set of all subjects. The performance of the trained models was compared in two aspects: overall recognition accuracy and individual recognition accuracy of each subject. From the experiment of evaluating the performance of the models according to the individual recognition accuracy of each subject, we only present the experimental results of the case in which the window length is set as 500 ms. The influence of our proposed TFFT suite on recognition accuracy was determined by comparing the results of the CNNs, LSTMs, TFFT-CNNs, and TFFT-LSTMs.

Figure 7 presents the overall recognition accuracy results of each model for different window lengths. When the window length was 200, 300, and 400 ms, the proposed TFFT-LSTMs obtained the best performance on NinaPro DB8. When the window length was 500 ms, the proposed TFFT-CNNs obtained the highest recognition accuracy. In general, the hybrid models based on the TFFT suite outperformed the pure models, demonstrating that performance was improved by introducing prior information to fine-tune the latent characteristics of the input sEMG signals.

Figure 8 presents the recognition accuracy of each evaluated model on different participants. The traditional methods based on CNN and LSTM obtained a higher recognition accuracy for participants $s4$ and $s6$ but a poor effect for some individuals, such as $s8$. The evaluated model based on our proposed TFFT suite not only improved the gesture recognition accuracy of all individuals but also narrowed the gap in the recognition performance in the data of different individuals.

https://doi.org/10.1017/S026357472200159X Published online by Cambridge University Press

Table I. *Training hyperparameter setting.*

Training Hyperparameters	Values
Epochs	120
Batch size	1200
Dropout rate	0.2
Initial learning rate	1e-3
Decayed factor of learning rate	10
Decayed epochs of learning rate	[60, 100]
Channels of backbone network	32
Channels of TFFT suite	16

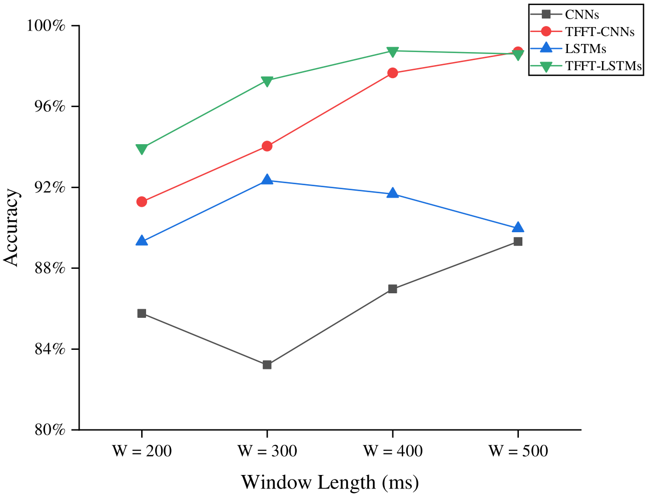


Figure 7. *Comparison of recognition accuracy of original architecture approaches.*

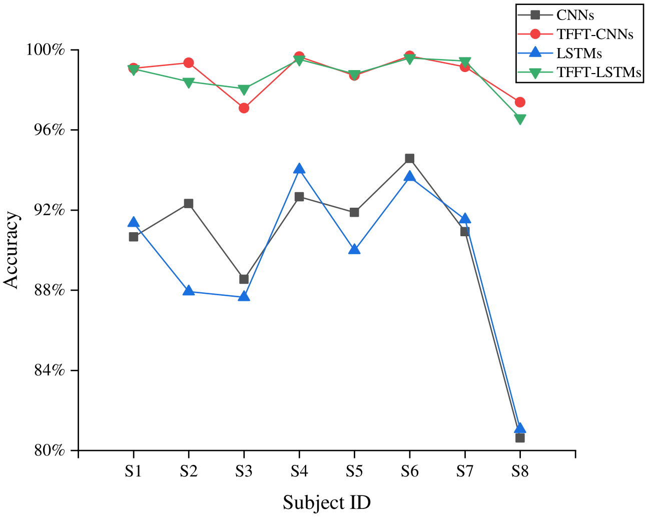


Figure 8. *Comparison of recognition accuracy of different subjects.*

https://doi.org/10.1017/S026357472200159X Published online by Cambridge University Press

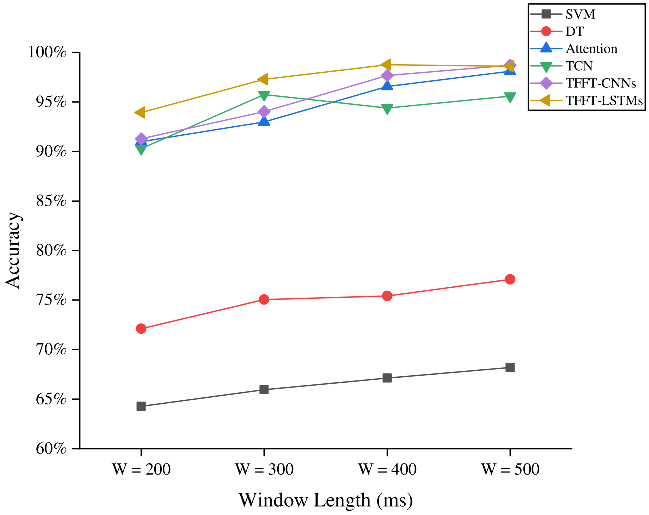


Figure 9. *Comparison of recognition accuracy of different state-of-the-art frameworks and classical algorithms.*

5.2.2. Comparison with different state-of-the-art frameworks and classical algorithms

To demonstrate the superiority of the TFFT suite, we compared the original TFFT architecture with existing state-of-the-art frameworks and classical algorithms using temporal features. Specifically, we chose two state-of-the-art DL models, namely the attention model and the TCN. Two other support vector machines (SVM) and decision trees (DT) that use variance and mean absolute value as temporal features were used as comparisons. As shown in Fig. 9, the model with the TFFT suite achieved the best accuracy for each window. SVM and DT algorithms based on temporal features do not perform as well as other methods. Although attention and temporal convolution process temporal data effectively, their interpretability is poor. This result not only shows that TFFT is effective but also shows that it intuitively considers the extracted feature information.

Although the performance of all models improved as the window size increased, different models require different costs. When the attention mechanism calculates the correlation matrix, the time and space complexity is squared, and the number of network layers of the TCN also increases logarithmically as the window becomes larger. However, the LSTM and CNN with TFFT are unaffected, making TFFT irreplaceable.

5.2.3. Comparison with different prior maps

We quantitatively compared our hybrid models that generate prior maps by CWT with other types of prior maps. As we aimed to discuss the efficiency of the prior maps obtained by CWT, we further considered the zero and randomly generated matrix for each individual as the prior inputs. Hence, by using the NinaPro DB8 for training, we obtained six evaluated models TFFT-CNNs (zeros), TFFT-CNNs (rand), TFFT-CNNs (CWT), TFFT-LSTMs (zeros), TFFT-LSTMs (rand), and TFFT-LSTMs (CWT).

Figure 10 presents the recognition results of each model in the datasets obtained under different window lengths. The proposed hybrid models based on the TFFT suite and CWT yielded better outputs. Naive prior input concatenation is insufficient for exerting the necessary condition for representing the characteristics of subject-specific sEMG signals. Thus, using the prior graph obtained by CWT as input in the proposed TFFT suite is a feasible attempt at achieving considerable performance in multi-individual gesture recognition using sEMG signals.

https://doi.org/10.1017/S026357472200159X Published online by Cambridge University Press

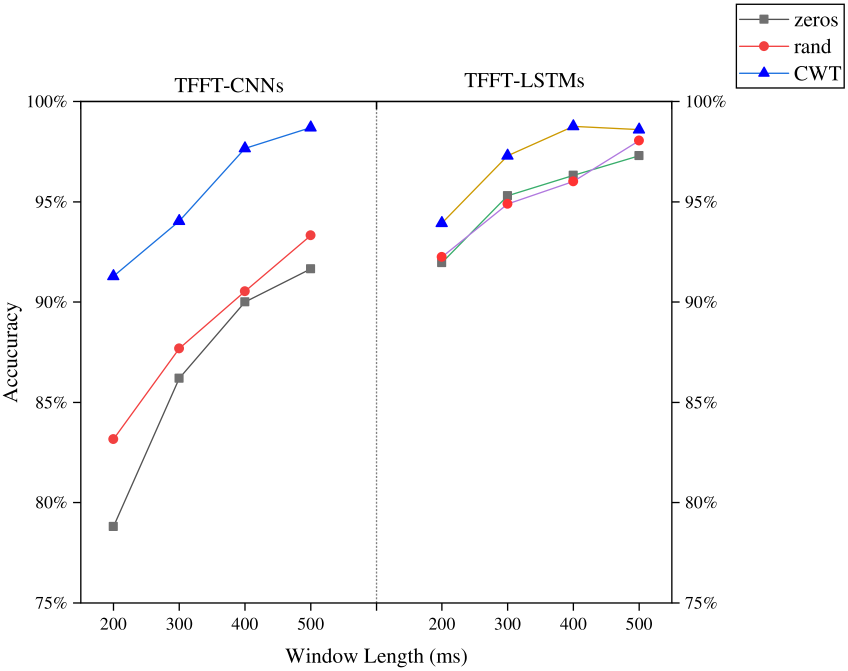


Figure 10. *Comparison of recognition accuracy of different priors.*

6. Conclusion

In this study, we investigated the use of a TFFT suite based on prior maps generated by CWT for multi-subject gesture recognition using sEMG signals. The advantages of TFFT are summarised as follows: (1) It enables the construction of a function with a few parameters that allow the intermediate feature transformation of a network in a single forward pass. (2) It can automatically capture necessary subject-specific information from prior input. (3) Each element of the intermediate features has independent affine transformation parameters. (4) It can be easily introduced into the existing network to construct an end-to-end network structure and the entire layers can be trained simultaneously.

The TFFT suite was introduced into CNNs and LSTMs and compared with pure CNNs and LSTMs, demonstrating superior recognition accuracy. The TFFT suite effectively narrowed the gap between the recognition performance in the data of different individuals. Moreover, LSTM and CNN equipped with TFFT outperformed two other state-of-the-art frameworks and classical algorithms, namely attention, TCN, SVM, and DT. We further demonstrated the feasibility of improving the recognition performance by introducing the maps generated using CWT in the sEMG signals from the human resting state as priors. However, this work is a first attempt at using CWT feature maps as prior knowledge and at designing the TFFT suite to learn modulation parameters. Therefore, we will perform different complex sEMG-based recognition tasks, consider different types of prior knowledge, and explore DL methods to enhance performance in the future work.

Author contribution. Can Wang and Xinyu Wu conceived the presented idea. Junpei Zhong developed the theory and performed the computations. Xin Zhou and Jiancong Ye verified the analytical methods. Can Wang encouraged Xin Zhou and Jiancong Ye to investigate gesture recognition and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

Financial support. This work was partly supported by International Science and Technology Cooperation Program of China (2018YFE0125600), National Natural Science Foundation-Shenzhen Joint Research Program (U2013213), and Shenzhen Fundamental Research project (JCYJ20210324120400003).

Conflicts of interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Balmik, A. Paikaray, M. Jha and A. Nandy, "Motion recognition using deep convolutional neural network for Kinect-based NAO teleoperation," *Robotica* **40**(9), 1–21 (2022).

[2] Y. Tao, Y. Huang, J. Zheng, J. Chen, Z. Zhang, Y. Guo and P. Li, "Multi-channel sEMG Based Human Lower Limb Motion Intention Recognition Method," **In:** *2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)* (IEEE, 2019) pp. 1037–1042.

[3] Y. Yuan, Z. Guo, C. Wang, S. Duan, L. Zhang and X. Wu, "Gait Phase Classification Based on SEMG Signals Using Long Short-Term Memory for Lower Limb Exoskeleton Robot," **In:** *IOP Conference Series: Materials Science and Engineering*. vol. **853** (IOP Publishing, 2020) pp. 012041.

[4] A. Gautam, M. Panwar, D. Biswas and A. Acharyya, "Myonet: A transfer-learning-based lrcn for lower limb movement recognition and knee joint angle prediction for remote monitoring of rehabilitation progress from sEMG," *IEEE J. Transl. Eng. Health Med*. **8**, 1–10 (2020).

[5] M. Atzori and H. Müller, "The Ninapro Database: A Resource for SEMG Naturally Controlled Robotic Hand Prosthetics," **In:** *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, 2015) pp. 7151–7154.

[6] N. Tsagkas, P. Tsinganos and A. Skodras, "On the Use of Deeper CNNs in Hand Gesture Recognition Based on sEMG Signals," **In:** *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (IEEE, 2019) pp. 1–4.

[7] Y. Hu, Y. Wong, W. Wei, Y. Du, M. Kankanhalli and W. Geng, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *PloS One* **13**(10), e0206049 (2018).

[8] Z. Ding, C. Yang, Z. Tian, C. Yi, Y. Fu and F. Jiang, "sEMG-based gesture recognition with convolution neural networks," *Sustainability* **10**(6), 1865 (2018).

[9] T. M. Bittibssi, A. H. Zekry, M. A. Genedy, S. A. Maged, "sEMG pattern recognition based on recurrent neural network," *Biomed. Sig. Proces.* **70**, 103048 (2021).

[10] X. Liu, K. N. Khan, Q. Farooq, Y. Hao and M. S. Arshad, "Obstacle avoidance through gesture recognition: Business advancement potential in robot navigation socio-technology," *Robotica* **37**(10), 1663–1676 (2019).

[11] F. Orabona, C. Castellini, B. Caputo, A. E. Fiorilla and G. Sandini, "Model Adaptation with Least-Squares SVM for Adaptive Hand Prosthetics," **In:** *2009 IEEE International Conference on Robotics and Automation* (IEEE, 2009) pp. 2897–2903.

[12] W. Ren, J. Pan, X. Cao and M.-H. Yang, "Video Deblurring Via Semantic Segmentation and Pixel-Wise Non-Linear Kernel," **In:** *Proceedings of the IEEE International Conference on Computer Vision* (2017) pp. 1077–1085.

[13] S. Zhu, R. Urtasun, S. Fidler, D. Lin and C. C. Loy, "Be Your Own Prada: Fashion Synthesis with Structural Coherence," **In:** *Proceedings of the IEEE International Conference on Computer Vision* (2017) pp. 1680–1688.

[14] K.-H. Park and S.-W. Lee, "Movement Intention Decoding Based on Deep Learning for Multiuser Myoelectric Interfaces," **In:** *2016 4th International Winter Conference on BRAIN-COMPUTER INTERFACE (BCI)* (IEEE, 2016) pp. 1–2.

[15] M. Atzori, M. Cognolato and H. Müller, "Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands," *Front. Neurorobot*. **10**, 9 (2016).

[16] N. Nasri, S. Orts-Escolano, F. Gomez-Donoso and M. Cazorla, "Inferring static hand poses from a low-cost non-intrusive semg sensor," *Sensors* **19**(2), 371 (2019).

[17] M. Simão, P. Neto and O. Gibaru, "EMG-based online classification of gestures with recurrent neural networks," *Pattern Recogn. Lett.* **128**, 45–51 (2019).

[18] A. Samadani, "Gated Recurrent Neural Networks for Emg-Based Hand Gesture Classification. a Comparative Study," **In:** *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, 2018) pp. 1–4.

[19] S. Shen, K. Gu, X.-R. Chen, M. Yang and R.-C. Wang, "Movements classification of multi-channel sEMG based on cnn and stacking ensemble learning," *IEEE Access* **7**, 137489–137500 (2019).

[20] H. Chen, Y. Zhang, G. Li, Y. Fang and H. Liu, "Surface electromyography feature extraction via convolutional neural network," *Int. J. Mach. Learn. Cyb.* **11**(1), 185–196 (2020).

[21] T. Matsubara and J. Morimoto, "Bilinear modeling of EMG signals to extract user-independent features for multiuser myoelectric interface," *IEEE Trans. Biomed. Eng.* **60**(8), 2205–2213 (2013).

[22] A. Xiong, X. Zhao, J. Han, G. Liu and Q. Ding, "An User-Independent Gesture Recognition Method Based on Semg Decomposition," **In:** *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2015) pp. 4185–4190.

[23] R. N. Khushaba, "Correlation analysis of electromyogram signals for multiuser myoelectric interfaces," *IEEE Trans. Neur. Syst. Rehab.* **22**(4), 745–755 (2014).

[24] B. Xue, L. Wu, K. Wang, X. Zhang, J. Cheng, X. Chen and X. Chen, "Multiuser gesture recognition using sEMG signals via canonical correlation analysis and optimal transport," *Comput. Biol. Med.* **130**, 104188 (2021).

[25] B. He, C. Wang, H. Wang, M. Li, S. Duan and X. Wu, "A Method for Recognition of Dynamic Hand Gestures Based on Wrist Tendon Sounds," **In:** *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)* (IEEE, 2021) pp. 1–6.

https://doi.org/10.1017/S026357472200159X Published online by Cambridge University Press

[26] M. Misiti, *Wavelet Toolbox for Use with MATLAB: User's Guide; Version 2; Computation, Visualization, Programming (MathWorks Incorporated*, 2000).

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is all you need," *Adv. Neur. Inform. Process. Syst.* **30** (2017).

[28] S. Bai, J. Z. Kolter and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, *arXiv preprint arXiv:1803.01271* (2018).

[29] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. M. Hager, S. Elsig, G. Giatsidis, F. Bassetto and H. Müller, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Sci. Data* **1**(1), 1–13 (2014).

[30] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization, arXiv preprint arXiv: 1412.6980 (2014).

Cite this article: X. Zhou, J. Ye, C. Wang, J. Zhong and X. Wu (2023). "Time–frequency feature transform suite for deep learning-based gesture recognition using sEMG signals", Robotica **41**, 775–788. https://doi.org/10.1017/S026357472200159X