Feasibility Analysis of sEMG Recognition via Channel-Wise Transformer

Jiaxuan Zhang†*, Yuki Matsuda†, Manato Fujimoto‡, Hirohiko Suwa†, and Keiichi Yasumoto†
† *Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan*
‡ *Osaka Metropolitan University, Osaka, Osaka 558-8585, Japan*

***Abstract*—Surface Electromyography (sEMG) features as objective biomarkers have the potential to facilitate a refined diagnosis, appropriate treatment/rehabilitation, and measurements of Human-Computer Interfaces (HCI). Precisely recognizing the informative features in EMG data is a staple challenge due to its massive, noise-involved, and nonstationary nature. In this paper, we build a channel-wise method assembling the recent powerful Transformer model for recognizing muscle activation. The gesture classification benchmark is demonstrated for proof-of-concept by introducing Ninapro sEMG datasets with the 17 gestures recorded from 40 subjects. From experiment results, our proposed method achieves a competitive overall accuracy of 85.6% by leveraging the sole channel-Transformer.**

I. Introduction

Studying neurophysiological processes is an important step toward understanding the control mechanism in human. Surface Electromyography (sEMG) as a major skeletal muscle imaging tool, exhibits the activation of the muscle cells with millisecond resolution [1]. The sEMG recognition is tough research typically, since the sEMG data is temporal-random, redundant, and noise-involved from the signal perspective [2]. Also, the spontaneous characteristics recorded in signals are transient without any pre-indicator [3]. Therefore, the above traits lead to the demands of precise methodologies in sEMG recognition for refining and distilling more important features.

Numerous studies make effort to yield reliable features for machine-learning in various sEMG-based applications. Conventionally, extraction of features is done by manually selecting/crafting features from the frequency or time-frequency domain. However, manual processing is laborious and practitioners seldom consider the feature extraction problem from a physiological viewpoint: the different channel of sEMG detects the electric potential generated by muscle cells when a specific muscle is electrically or neurologically activated.

It has been demonstrated that combining with the automatic feature extraction using deep neural network, especially convolutional neural networks (CNNs) could improve the feature mapping and the downstream task performance [4]. Those works typically leveraged a set of CNNs as filters to capture more informative local features from sEMG recordings and generate several channel-wise feature maps. However, though the convolution operation can express the channel connectivity of different muscle signals, the local inductive bias of CNNs leads to the loss of global context. Different from the CNNs, the recurrent neural networks (RNNs) pay attention to the information of global context by allowing sequential modeling of dependency and transfer of temporal influence. However, while periodic components of sEMG can be captured by RNNs, other components such as transiently burst rhythms activated by muscle cells are generally unpredictable. Further, the inherently sequential attribute (time-invariant) of RNNs precludes the possibility of parallelization in feature capture.

Considering the parallel computation in finding the specific activated muscle and the corresponding sensing channel, this study introduces recent attention-mechanism-based model, i.e., the Transformer to sEMG recognition. Specifically, we follow the most of existing sEMG-based works to extract the time-frequency domain features to represent each channel information of sEMG, and then adopt the Transformer to importance of channels in parallel. For proof-of-concept, we experiment a proposal with the gesture classification problem and the results are promising comparing with the existing works.

This work was partly supported by JSPS KAKENHI JP21K19828.
* Corresponding author (e-mail: zhang.jiaxuan.ze4@is.naist.jp)

II. Method

A. Dataset&Preprocessing

In this study, we evaluated the proposal on the second Ninapro benchmark dataset (termed by DB2) [5] for 17 gesture recognition. The DB2 consists of 40 healthy subjects while each subject collects the 12 channels sEMG signal with the 2 kHz sample rate. A Butterworth bandpass (5-500 kHz) filter within 3-order is applied to each subject signal. Then a 200 msec moving window with half overlapping is utilized for segmentation and sample generation by referring to [2]. Hence, a total of 81,732 sEMG samples are used in this work.

B. Time-Frequency Feature Extraction

We adopt the dual-tree complex transformation, to each raw data that transform the time-domain to time-frequency domain signal. Then, a feature matrix is generated by following six classical sEMG feature extraction methods for each sample.

- Mean absolute value
- Standard deviation
- Zero crossing
- Average energy
- Waveform length
- Max fractal length

In summary, each row of the feature matrix illuminates one channel feature vector of sample, and the initial feature is denoted as $x \in \mathbb{R}^{C \times F}$, where C denotes the channels (12) and F denotes the six classical features.

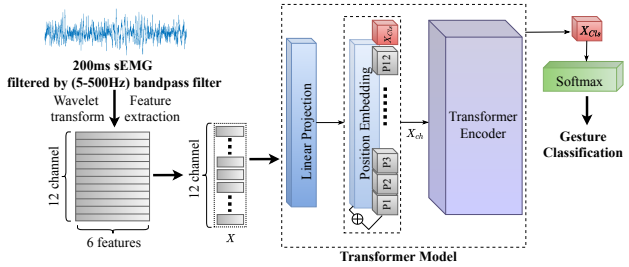
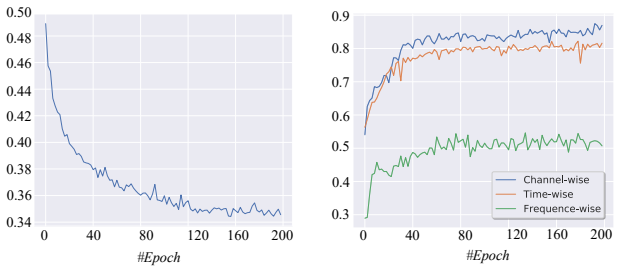


Fig. 1. Proposal overview



(a) Training loss (b) Validation comparison

Fig. 2. (a) Training loss in channel-wise Transformer; (b) Accuracy in different ablation studies (time/frequency-wise Transformer comparing to ours)

C. Channel-Wise Transformer

In this study, we assume the Transformer [6] is an appropriate candidate that can distill the important channel for the precise classification of each gesture. Therefore, we view each row (channel) of matrix x as one feature *patch* (well-known embedding operation) for the input of the Transformer where each patch is demonstrated by extracted six statistical features from the wavelet signal. Then a *patch-wise* linear layer $\sigma(\cdot)$ expands the dimension of each *patch* to a more informative space. An extra class token x_{Cls} is inserted to each *patch* sequence at the beginning and used for the final classification decision [7]. The above steps can be formulated as below:

$$x_{ch} = \texttt{Concat}(x_{Cls}, \sigma(x)), \tag{1}$$

where $x_{seq} \in \mathbb{R}^{(12+1)\times D}$ is the output sequence, where D is the dimension of the linear layer output and we set to 32, here.

The *patches* pass through the network architecture as shown in Fig. 1. The attention network associates the individual patches and maps the relevance to the gesture label y_i,$i \in \mathbb{R}^{17}$ with three components: the query (Q), key (K), and value (V) matrices, which are the projected matrices of the input X_{ch}. Here, K is the same as Q, and the attention is utilized to calculate relevance among the *patches*. The resultant relevance values can be viewed as a weight-set and are further used to V.

$$\texttt{Attention} = \texttt{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \tag{2}$$

where operator $\sqrt{d}$ denotes a normalization. When the input x_{ch} passes through the Transformer, the $CrossEntropy$ is applied to calculate the loss between x_{Cls} and y_i and minimize the resultant error.

III. Experiment&Result

Fig. 2 (a) exhibits that our method does not trap into the overfitting issue due to the smooth convergence trend during the training phase. The ablation studies in Fig. 2 (b) prove the effectiveness of the proposed method which analyzes the sEMG signals by paying attention to the specific muscle activation. Table I presents a comparison of the overall accuracy (5-fold cross validation) for our proposal and two existing works. The proposed Transformer has a better performance than the CNNs, even when we solely extract six classical features. Comparing to the state-of-the-art work [2], our proposal also got a competitive accuracy (i.e., 85.6% which is comparable to 85.9% [2]) without complicated CNNs filters. Such results prove the feasibility in solely modeling the channel information by the Transformer model.

TABLE I
Comparison of performance among two related works and our pipelines for DB2

	Method	Subject	Accuracy
TBME, 2019 [1]	CNNs	40	83.7%
TNSRE, 2021 [2]	CNNs + Temporal Attention	40	85.9%
Our proposal	Channel-wise Attention	40	85.6%

IV. Conclusion

This paper presented a channel-wise Transformer model for recognizing gestures from sEMG signals. We utilized the time-frequency features to represent the channel of sEMG, and introduced the attention-mechanism to distill the important channels. The superior results shed a light on how one should extract and recognize sEMG features. Interesting future directions include investigating the available application in more fundamental biomedical problems, such as the gait detection.

References

[1] W. Wei, Q. Dai, Y. Wong, Y. Hu, M. Kankanhalli, and W. Geng, "Surface-electromyography-based gesture recognition by multi-view deep learning," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2964–2973, 2019.

[2] E. Rahimian, S. Zabihi, A. Asif, D. Farina, S. F. Atashzar, and A. Mohammadi, "Fs-hgr: Few-shot learning for hand gesture recognition via electromyography," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1004–1015, 2021.

[3] Z. Chen, L. Zhu, Z. Yang, and R. Zhang, "Multi-tier platform for cognizing massive electroencephalogram," in *arXiv*, 2022.

[4] Z. Chen, N. Ono, W. Chen, T. Tamura, M. Altaf-Ul-Amin, S. Kanaya, and M. Huang, "The feasibility of predicting impending malignant ventricular arrhythmias by using nonlinear features of short heartbeat intervals," *Computer Methods and Programs in Biomedicine*, vol. 205, p. 106102, 2021.

[5] S. Pancholi, P. Jain, A. Varghese, and A. M. joshi, "A novel time-domain based feature for emg-pr prosthetic and rehabilitation application," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 5084–5087.

[6] Z. Chen, Z. Yang, D. Wang, M. Huang, N. Ono, M. Altaf-Ul-Amin, and S. Kanaya, "An end-to-end sleep staging simulator based on mixed deep neural networks," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 848–853.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.