郭珺, 孙皖京, 李素姣. 基于 Transformer 的实时手部运动识别框架[J]. 智能计算机与应用,2025,15(11):142-147. DOI: 10.20169/j.issn.2095-2163.251123

基于 Transformer 的实时手部运动识别框架

郭 珺[1], 孙皖京[1], 李素姣[1,2,3]

(1 上海理工大学 健康科学与工程学院, 上海 200093; 2 上海康复器械工程技术研究中心, 上海 200093;
3 民政部神经功能信息与康复工程重点实验室, 上海 200093)

摘 要: 基于表面肌电信号(Surface Electromyography,sEMG)的假肢动作识别方法因其良好的应用前景成为智能假肢控制领域的重要研究方向。为解决传统 CNN 和 LSTM 在训练效率和实时响应上的局限性,本文将 Transformer 架构引入实时手部肌电信号识别任务,构建了一个基于 Transformer 的实时手部运动识别框架,利用 Transformer 特有的时空特征捕获能力,提高实时手部肌电信号识别性能。通过对比实验,验证了其在识别性能、训练效率和实时性等方面相较于 CNN 和 LSTM 模型的优势。

关键词: 智能假肢; 手部运动识别; 表面肌电信号

中图分类号: TP242.6 **文献标志码**: A **文章编号**: 2095-2163(2025)11-0142-06

Transformer-based framework for real-time hand motion recognition

GUO Jun[1], SUN Wanjing[1], LI Sujiao[1,2,3]

(1 School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
2 Shanghai Engineering Research Center of Rehabilitation Devices, Shanghai 200093, China;
3 Key Laboratory of Neural Function Information and Rehabilitation Engineering, Ministry of Civil Affairs, Shanghai 200093, China)

Abstract: Surface electromyography (sEMG)-based prosthetic gesture recognition has emerged as a significant research focus in intelligent prosthetic control due to its promising application prospects. To address the limitations of traditional Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks in training efficiency and real-time performance, this study introduces the Transformer architecture into the task of real-time hand motion recognition using sEMG signals, developing a transformer-based real-time hand motion recognition framework. This framework aims to leverage the transformer's distinctive capability for capturing spatiotemporal features to enhance the performance of real-time sEMG-based hand motion recognition. Through comparative experiments, this paper demonstrates the framework's superiority over CNN and LSTM models in terms of recognition accuracy, training efficiency, and real-time performance.

Key words: intelligent prosthetics; hand motion recognition; sEMG

0 引 言

研发具备高度仿生功能及感知功能的理想假肢,对于提升截肢者的生活自理能力及社会参与度具有重要价值。

在智能假肢的研究中,基于 sEMG 控制的假肢因其能够最大限度地还原人类上肢功能而成为研究的热点[1]。随着模式识别技术在肌电假肢控制接口中的应用,假肢的控制可以实现更加直接和多自由度的动作[2]。卷积神经网络(CNN)和长短期记忆网络(LSTM)作为主流深度学习模型,在基于 sEMG 信号的肌电假肢模式识别任务中得到广泛应用[3]。通过设计改良的 CNN 架构,可以提升分类效能,Geng 等[4]提出将原始 sEMG 信号编码为时频图像,输入至 CNN 中进行分析,增强空间特征提取能力;Zhai 等[5]采用时延叠加频谱图丰富特征表示;Wei 等[6]设计多级分解与融合策略训练 CNN 模型,优化识别能力。为更好地捕捉长期依赖关系,克服梯度消失问题,引入了长短时记忆网络(LSTM)模型。Bao 等[7]构建端到端 CNN-LSTM 混合模型,

基金项目: 国家重点研发计划项目(2020YFC2007902)。

作者简介: 郭 珺(1999—),女,硕士,主要研究方向:肌电假肢,模式识别; 孙皖京(1999—)男,硕士,主要研究方向:肌电假肢,模式识别。

通信作者: 李素姣(1979—),女,博士,副教授,主要研究方向:智能假肢,康复机器人。Email: sujiao.li@usst.edu.cn。

收稿日期: 2024-03-05

CNN 提取空间特征后由 LSTM 捕获动态时序模式；Huang 等[8]将 sEMG 信号编码为时频谱图，输入 CNN-LSTM 双分支网络，实现时空联合建模；Bai 等[9]引入通道注意力机制，优化多通道 sEMG 的 CNN-LSTM 特征融合权重，显著提升了分类准确率与模型鲁棒性。

近年来，Transformer 架构在自然语言处理与计算机视觉领域取得进展，其应用已扩展至 sEMG 动作识别研究。Montazerin 等[10]建立 CT-HGR（Compact Transformer-based Hand Gesture Recognition）框架，通过高密度 sEMG 空间拓扑编码，提升跨主体识别鲁棒性；Montazerin 等[11]采用 Transformer 模型，降低底层模型复杂性，解决模型训练时间问题；Godoy 等[12]设计多通道时序视觉 Transformer，优化灵巧操作动作的注意力权重分配。相较于 CNN 和 LSTM，Transformer 通过全局自注意力建立序列元素直接关联，在减少参数量的同时提升训练效率。

然而，现有研究尚未充分发挥 Transformer 在实时肌电信号处理中的潜力。本文提出一种面向手部肌电信号实时识别的时序注意力 Transformer 框架（Temporal Attentional Transformer for sEMG Gesture Recognition，TATEGR）。该框架充分利用 Transformer 的高效时空特征捕获能力，旨在解决传统方法在训练效率和实时响应方面的不足。通过对比实验验证其相对于传统 CNN 和 LSTM 模型在性能、训练效率和实时性等方面的优越性。

1　基于 Transformer 的实时手部运动识别框架

1.1　基于 Transformer 基础架构与技术原理

Transformer 是 Vaswani 等[13]2017 年提出的一种深度学习模型，其在自然语言处理（NLP）领域实现了突破性进展。该模型摒弃了传统的循环神经网络（RNN）和卷积神经网络（CNN）以多头自注意力机制（Multi-Head Self-Attention）为核心构建模型，使得模型能够对输入序列中的每个位置的信息进行全局建模，并且各个位置之间可以相互依赖、互相影响，解决了 RNN 在长序列处理时存在的梯度消失或爆炸问题。自注意力权重的计算过程如下：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)V \quad (1)$$

其中，$\boldsymbol{Q}$ 表示查询矩阵，用于衡量序列中每个位置的元素与其他所有位置元素的相关性；$\boldsymbol{K}$ 表示键矩阵，用于存储每个位置的关键信息；$\boldsymbol{V}$ 表示值矩阵，包含了每个位置处的有用信息价值；$\boldsymbol{K}^{\mathrm{T}}$ 是键矩阵的转置，与查询矩阵相乘，计算查询向量与所有键向量的内积，体现了序列内部的位置关联度；d_k 是向量维度，除以 $\sqrt{d_k}$ 是为了归一化，防止由于向量维度 d_k 增大导致的注意力得分过大，保证了 Softmax 函数的稳定性。

除此之外，Transformer 还引入了位置编码（Positional Encoding）来保留序列信息，以及前馈神经网络（Feed Forward Network，FNN）作为其全连接层部分，进一步增强模型的表达能力。这种分层并行结构设计极大地提升了计算效率，尤其适合 GPU 等硬件加速设备进行高效训练。

1.2　肌电信号处理

本文数据预处理主要包括滤波降噪、标准化和活动段提取。根据 sEMG 信号的频谱能量分布特性，使用三阶 10 Hz 巴特沃斯高通滤波器去除了电缆带来的运动伪影和电气干扰。

由于不同受试者的解剖结构和生理条件各异，多通道 sEMG 信号呈现显著的个体差异性。为了减弱这类差异对后续模式识别和分类精度的影响，本文采用 Z-Score 标准化方法对各受试者的 sEMG 数据进行了规范化处理，公式如下：

$$Z = (X - \mu)/\sigma \quad (2)$$

其中，X 代表原始信号值；μ 为信号的均值；σ 为信号的标准偏差。

该方法将不同量纲、不同大小的数据转换为无量纲、零均值、单位方差的标准分数形式，确保跨受试者数据的一致性。为精准识别出 sEMG 信号中的实质性活动片段，采用了一种自适应双阈值提取策略代替传统的单一阈值识别方法。本文设定一个动态阈值系统，能够在不稳定的肌电信号环境中准确捕捉到肌肉活动的有效时段计算方法。

（1）对单通道肌电信号差分处理，生成瞬时平均能量序列 E：

$$E = \frac{1}{N}\sum_{i=1}^{N}[s_k(i+1) - s_k(i)]^2 \quad (3)$$

其中，N 代表通道总数，本实验 $N=8$，$s_k(i)$ 代表第 i 通道在时刻 k 的采样值。

（2）计算 E 在窗口长度为 64 ms 内的能量均值 S：

$$S = \frac{1}{L}\sum_{j=1}^{j+L-1}E(j) \quad (4)$$

其中，L 代表滑动窗口时长，本实验 $L=64$，

$E(j)$ 表示由式(3)计算的时刻 j 的瞬时能量。

(3)基于 S 的中位数以及方差生成动态阈值 Th_1 和 Th_2:

$$\mathrm{Th}_1 = S, \quad 0 < c < \mathrm{Var}(S) \tag{5}$$

$$\mathrm{Th}_2 = \mathrm{Median}(S), \quad c < 0 \tag{6}$$

其中,$\mathrm{Median}(S)$ 代表能量均值 S 的中位数,$\mathrm{Var}(S)$ 代表序列 S 的方差,$c = S - \mathrm{Median}(S)$ 。

1.3 用于肌电信号手部运动识别的 Transformer 框架

Transformer 架构虽源于自然语言处理领域,但其核心的自注意力(Self-Attention)机制,因其强大的序列建模能力,已被广泛应用于各类序列数据处理任务,包括生物医学信号分析,如基于表面肌电信号的手势识别。在基于 sEMG 的手势识别任务中,Transformer 架构的优势在于其对信号时空特征的全局建模能力和动态特征提取能力。

本文设计用于肌电信号手势识别的时序注意力 Transformer 框架(TATEGR)如图 1 所示。首先,对采集的肌电信号进行标准的数据预处理,包括但不限于噪声去除、滤波处理、适当的时间分割以及特征抽取,最终将原始生物电信号转化为符合模型输入需求的时间序列数据集;其次,将预处理后的肌电信号表示为一系列连续的向量序列。在 TATEGR 框架下,每一个时间步长的肌电特征均被映射为 $L \times N$ 的向量表示,构成输入序列,L 代表每个动作的肌电信号序列长度,N 代表肌电信号采集通道数。本文中 L =4 096,N = 8。

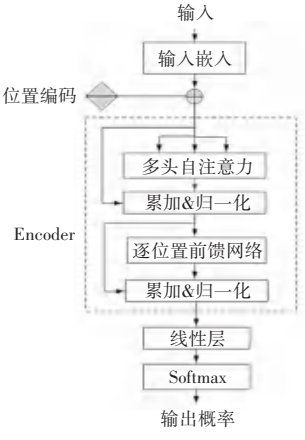


图 1 用于肌电信号手势识别的时序注意力 Transformer 框架

Fig. 1 Temporal attentional Transformer for sEMG gesture recognition

TATEGR 将输入大小为 $x \in R^{L\times N}$ 的散斑图像分成维度为 $A \times A$ 的 M 个片段向量,即 $x_p \in R^{(M\times(A^2\cdot N))}$,通过位置编码模块将这些片段向量映射到一个 $\boldsymbol{W}$ 维的嵌入空间中,并引入一个特定的位置嵌入向量 $\boldsymbol{E}_{\mathrm{pos}}$,该向量与每个片段向量相对应,旨在编码并保留其原有的空间信息,从而形成如下式所示的向量 $\boldsymbol{w}$:

$$\boldsymbol{w} = [x_p^1 E, x_p^2 E, \cdots, x_p^N E] + \boldsymbol{E}_{\mathrm{pos}} \tag{7}$$

其中,E 表示在每个片段向量间嵌入的投影。

TATEGR 的核心组成部分——编码模块,对接收到的向量 $\boldsymbol{w}$ 进行深入处理。编码模块通过多头自注意力机制(MSA)和逐位置前馈网络(Position-wise Feed-Forward Networks,FFN)来学习向量 $\boldsymbol{w}$ 的全局上下文表示。MSA 通过 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 3 组矩阵,让模型能够对输入序列中的任意两个位置进行比较并学习两者之间的关联性。FFN 对输入向量进行非线性转换,并通过进一步提炼和丰富每个位置的特征表达,增强了模型捕捉复杂依赖关系的能力。编码模块中第 i 个 MSA 层和 FFN 层的输入结果分别如下式:

$$w_i' = \mathrm{MSA}(\mathrm{Norm}(w_{i-1})) + w_{i-1} \tag{8}$$

$$w_i = \mathrm{FFN}(\mathrm{Norm}(w_i')) + w_i' \tag{9}$$

其中,Norm 表示归一化层;w_{i-1} 为第 i 个 MSA 层输入的结果;w_i 为第 i 个 FFN 层输出的结果。

在 Encoder 编码之后,模型将编码结果依次输入线性层和归一化层,以将 Encoder 输出的高层次抽象特征映射至预定义的手势类别空间,生成各类别的概率分布估计,并最终判断手势动作类别。分类结果输出如下:

$$\hat{c} = \mathrm{argmax}\ (\mathrm{Softmax}\ (\boldsymbol{W}\boldsymbol{w} + \boldsymbol{b})) \tag{10}$$

其中,$\boldsymbol{W}$ 是线性层的权重矩阵;$\boldsymbol{b}$ 是偏置向量;$\hat{c} \in C$ 为最终判断的手势动作类别。

最后,此模型借助已标记的肌电信号手势数据集进行训练优化,利用反向传播算法调整模型参数,提高手势模式识别的准确性。

针对 sEMG 信号的动作分类任务中,首先通过位置编码赋予信号的时间属性,随后运用自注意力层来探索不同通道间及同一通道不同时刻信号间的交互作用,有效识别关键肌肉活动区域和时段。此外,TATEGR 完全并行化的结构特点在大规模肌电信号数据集上的训练更为高效,有利于提高模式识别的准确率和实时性能。

2 实验结果与分析

2.1 实验方案设计

皮肤表面肌电信号采集设备采用上海傲意生产

的 gForcePro+肌电专业臂环，该臂环有 8 个肌电采集通道，采样频率为 1 000 Hz。实验采集 20 名受试者 10 种手部运动模式下的 sEMG 信号，手部运动模式如图 2 所示。受试者根据屏幕提示做对应的动作，每个动作保持 3 s，休息 3 s，重复 8 次，不同动作休息间隔时间 3 min。受试者尽可能快速、自然且一致的完成手势任务。在系统提示休息时，受试者放松手臂肌肉。

图 2　手部运动模式图

Fig. 2　Hand movement pattern diagram

实验设计方案如图 3 所示。在采集并预处理 20 名受试者的数据之后，采取 64 ms 窗口长度和 64 ms 滑动步长来提取和组织数据，形成离线实验数据集，将数据集以 8∶2 的比例分为训练数据与测试数据，输入本文设计的 TATEGR 分类模型以及用于对照的 CNN 和 LSTM 模型。通过该离线实验数据集对应的识别性能和混淆矩阵，检验 TATEGR 模型的分类效能。同时，通过个性化模型在线识别所展示的识别性能和响应速度，验证了 TATEGR 模型在实时识别效能和实时性。

所有参与对比的模型在训练时保持一致的超参数配置：设置训练轮次为 30 Epoch，批量大小(batch size)固定为 128，采用 Adam 优化算法，并初始化学习率为0.001。此外，均统一使用交叉熵损失函数作为模型训练过程中的损失函数，以此来度量和比较各个模型的分类性能。

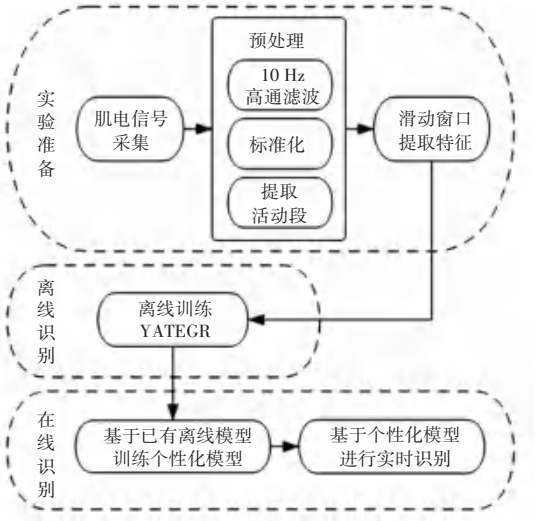


图 3　实验设计方案

Fig. 3　Experimental design scheme

2.2　离线实验结果分析

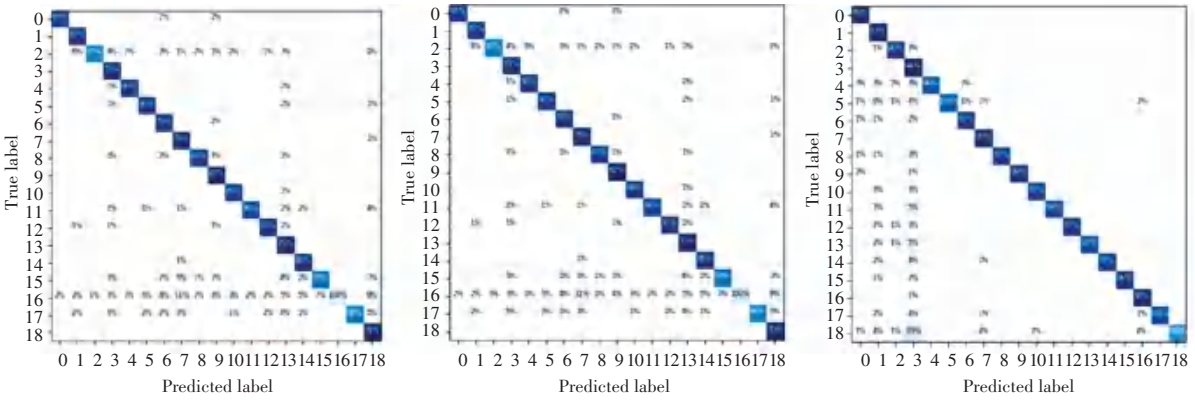
CNN，LSTM 和 TATEGR 3 种模型的离线识别实验结果见表 1。TATEGR 模型在测试集上的准确率达到 95.99%，远超 CNN 模型的 78.42%，同时也略高于 LSTM 模型的 92.01%，表明 TATEGR 在处理复杂多运动模式识别任务时，具备更高的判别准确性；其次，在召回率、精确率和 *F*1 分数这 3 个衡量分类效果的重要指标上，TATEGR 同样得到了最好的结果，分别为 95.56%、96.54% 和 96.05%，证明了 TATEGR 在各类运动模式的识别上具有高覆盖率、低误报率以及良好的综合性能；最后，在训练效率方面，TATEGR 模型仅需 401 min 即可完成训练，不仅显著优于 LSTM 模型，且比 CNN 模型更快，充分展现了 TATEGR 在训练效率方面的显著优势，使得 TATEGR 在手势识别的大型数据集训练或者需要快速迭代优化模型的场景中，展现出更强的竞争优势和实用性。

表 1　3 种不同模型离线识别结果

Table 1　Offline identification results of three different models

模型	数据集	损失值	召回率/%	准确率/%	精确率/%	*F*1	训练时间/min
CNN	训练集	0.473 7	78.57	83.21	88.57	0.832 7	522
CNN	测试集	0.644 9	73.72	78.42	84.76	0.785 9	N/A
LSTM	训练集	0.233 9	91.30	92.01	93.17	0.922 3	2 583
LSTM	测试集	0.390 5	86.27	86.53	83.55	0.848 9	N/A
TATEGR	训练集	0.115 5	95.56	95.99	96.54	0.960 5	401
TATEGR	测试集	0.326 9	85.46	88.76	91.29	0.882 8	N/A

3 种模型识别结果混淆矩阵如图 4 所示。CNN 模型在识别 9 个特定动作时准确率低于 90%,仅 3 个动作识别准确率超过 95%;LSTM 模型则有 4 个动作识别率高于 95%,8 个动作低于 90%。相比之下,TATEGR 模型整体识别性能最优,除动作 1(伸拇指)和动作 3(伸食指)外,其余动作识别准确率均超过 90%,其中 8 个动作甚至达到 99%以上,但动作 3 的识别率仅为 44%,可能是模型出现了一定程度的过拟合现象。



(a) CNN 混淆矩阵 (b) LSTM 混淆矩阵 (c) TATEGR 混淆矩阵

图 4 3 种模型识别结果混淆矩阵

Fig. 4 Confusion matrices of the recognition results of three models

TATEGR 训练集各评价指标见表 2。由表 2 可知,所有受试者离线准确率均达到了 97%以上,在召回率和精确率稳定的同时也呈现出较高水平。该数据表现与 TATEGR 离线识别结果较为一致,有效验证了使用新增受试者数据以更新并训练已有模型的可行性,而且减少了离线训练的时间,提高了整体训练效率。

表 2 TATEGR 训练集各评价指标结果

Table 2 Training set performance metrics results

实验者编号	损失值	准确率/%	召回率/%	精确率/%	*F*1
1	0.047 7	99.67	99.56	99.78	0.996 7
2	0.176 0	97.45	94.79	98.73	0.967 2
3	0.071 4	99.44	98.66	99.66	0.994 6
4	0.087 2	98.57	97.59	99.11	0.983 4
5	0.028 5	99.89	100.00	99.89	0.999 4
6	0.067 8	98.90	98.57	99.11	0.988 4
7	0.064 8	99.12	93.36	99.45	0.989 0
8	0.045 1	99.89	99.89	100.00	0.999 5
9	0.083 7	99.22	97.66	100.00	0.988 1
10	0.164 3	97.97	95.50	99.72	0.970 8
11	0.053 3	99.01	98.91	99.12	0.984 9
12	0.111 2	98.79	96.91	99.10	0.979 9
平均值	0.083 4	98.99	98.00	99.39	0.986 8

2.3 在线实验结果分析

12 位受试者各手势类别的平均准确率及其标准差如图 5 所示。由图 5 可知,12 位受试者的各动作的平均准确率为 93.74%,除伸食指和中指(动作 5),五指伸展(动作 14),五指捏(动作 15)外,其他动作模式的实时识别率均达到了 90%以上。对于这 3 种动作准确率相对较低的现象,推测其原因在于以上动作复杂的指尖伸展运动,尤其是涉及到拇指的活动,由于拇指的运动控制涉及的手部肌肉更为丰富,故此在表面肌电图(sEMG)信号中,这些动作产生的信号强度较大且相似性较高,从而使得其它手指细微动作的肌电信号特征可能被掩盖,降低了识别系统的分辨能力,即当 sEMG 信号提供的动作差异化特征不够鲜明时,特别是在涉及手指复杂伸展动作的情况下,运动模式的识别误差率可能会显著增加。

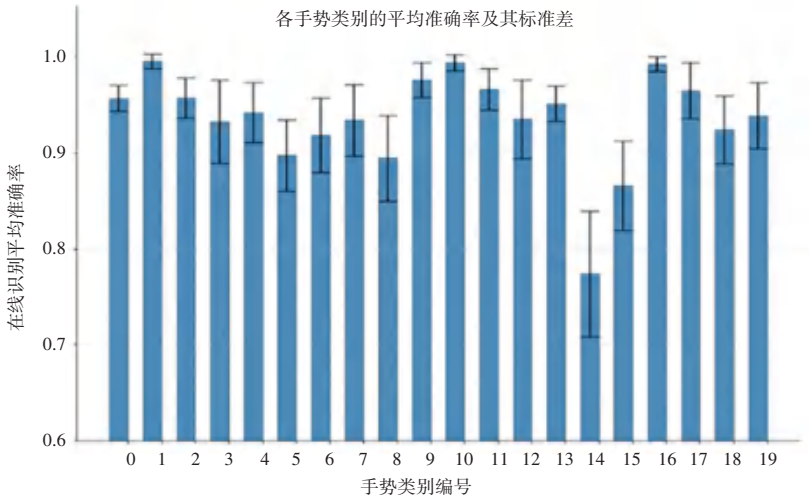


图 5　各手势类别的平均准确率及其标准差

Fig. 5　Average accuracy and standard deviation for each gesture category

3　结束语

本文针对肌电假肢实时控制需求，构建了一种用于肌电信号手势识别的时序注意力 TRANSFORMER 框架(TATEGR)，以提升智能假肢的功能性和实时性。在对既有 CNN 和 LSTM 模型的深入探讨和借鉴基础上，TATEGR 通过全局自注意力机制有效解决了时序依赖限制，降低了模型复杂度并提高了训练效率。实验结果证实，在离线模式识别场景中，TATEGR 识别准确率达到 95.99%，各项评价指标显著优于 CNN 和 LSTM 模型，并对各类动作的识别能力有了实质性增长。在线实时识别测试中，TATEGR 模型平均识别准确率显著提升，且平均时延降至 145 ms，展现出在肌电假肢运动模式识别中较强的实时性和低延迟特性，具有一定的临床应用和社会价值。

参考文献

[1] SAMUEL O W, ASOGBON M G, GENG Y, et al. Intelligent EMG pattern recognition control method for upper - limb multifunctional prostheses: advances, current challenges, and future prospects[J]. IEEE Access, 2019, 7: 10150-10165.

[2] XIONG D, ZHANG D, ZHAO X, et al. Deep learning for EMG -based human - machine interaction: A review[J]. IEEE/CAA Journal of Automatica Sinica, 2021, 8(3): 512-533.

[3] RANI G J, HASHMI M F, GUPTA A. Surface electromyography and artificial intelligence for human activity recognition: A systematic review on methods, emerging trends applications, challenges, and future implementation[J]. IEEE Access, 2023, 11: 105140-105169.

[4] GENG W, DU Y, XI W, et al. Gesture recognition by instantaneous surface EMG images[J]. Scientific Reports, 2016, 6(1): 36571.

[5] ZHAI X, JELFS B, CHANR H M, et al. Self - recalibrating surface EMG pattern recognition for neuroprosthesis control based on convolutional neural network[J]. Frontiers in Neuroscience, 2017, 11: 379.

[6] WEI W, WENG Y, DU Y, et al. A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface[J]. Pattern Recognition Letters, 2019, 119: 131-138.

[7] BAO T, ZAIDI S A R, XIE S, et al. A CNN-LSTM hybrid model for wrist kinematics estimation using surface electromyography[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-9.

[8] HUANG D, CHEN B. Surface EMG decoding for hand gestures based on spectrogram and CNN-LSTM[C]//Proceedings of 2019 2nd China Symposium on Cognitive Computing and Hybrid Intelligence (CCHI). Piscataway, NJ: IEEE, 2019: 123-126.

[9] BAI D, LIU T, HAN X, et al. Multi-channel sEMG signal gesture recognition based on improved CNN-LSTM hybrid models [C]//Proceedings of 2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR). Piscataway, NJ: IEEE, 2021: 111-116.

[10] MONTAZERIN M, RAHIMIAN E, NADERKHANI F, et al. Transformer-based hand gesture recognition from instantaneous to fused neural decomposition of high-density EMG signals[J]. Scientific Reports, 2023, 13(1): 11000.

[11] MONTAZERIN M, ZABIHI S, RAHIMIAN E, et al. ViT-HGR: Vision transformer-based hand gesture recognition from high density surface EMG signals[C]// Proceedings of 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Piscataway, NJ: IEEE, 2022: 5115-5119.

[12] GODOY R V, DWIVEDI A, LIAROKAPIS M. Electromyography based decoding of dexterous, in-hand manipulation motions with temporal multichannel vision transformers[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022,30:2207-2216.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems. NIPS, 2017: 5998-6008.