DOI:10.3969/j.issn.1003-5060.2024.05.002

一种双路并行的大规模手势识别模型

曹一丹，　王青山，　王　琦
(合肥工业大学 数学学院,安徽 合肥　230601)

摘　要:文章以大规模手势为研究对象,提出一种基于肌电信号(electromyography,EMG)分支和惯性测量单元(inertial measurement unit,IMU)分支的双路并行手势识别模型。首先,设计双路并行模型来充分提取数据特征,EMG 分支利用二维卷积神经网络设计双流结构,分别关注 EMG 信号的空间和通道变化,IMU 分支在卷积长短时记忆(convolutional long short-term memory,ConvLSTM)网络基础上引入时间机制,将空间信息与时间信息融合;其次,对模型预训练并根据预训练模型进行参数微调,提高模型泛化性;最后,在 500 个常用的中国手语手势上进行测试,结果表明,该模型平均识别率为 82.1%,与 SignSpeaker 和 CG-Recognizer 相比分别提高了 21.0%和 6.8%。

关键词:预训练;手势识别;深度学习;肌电信号(EMG);惯性测量单元(IMU)
中图分类号:TP183　　**文献标志码**:A　　**文章编号**:1003-5060(2024)05-0585-06

A large-scale gesture recognition model with dual-path parallel

CAO Yidan,　WANG Qingshan,　WANG Qi
(School of Mathematics, Hefei University of Technology, Hefei 230601, China)

Abstract: In this paper, a dual-path parallel gesture recognition model based on the electromyography (EMG) branch and the inertial measurement unit(IMU) branch is proposed for large-scale gestures. Firstly, the dual-path parallel model is designed to fully extract the data features. The EMG branch uses a two-dimensional convolutional neural network to design a dual-stream structure to focus on the spatial and channel variations of EMG signals, respectively. The IMU branch introduces a temporal mechanism based on the convolutional long short-term memory(ConvLSTM) network to fuse spatial and temporal information. Secondly, the model is pre-trained and the parameters are fine-tuned according to the pre-trained model to improve the generalization of the model. Finally, the model is tested on 500 commonly used Chinese sign language gestures, and the average recognition rate of the model is 82.1%, which is 21.0% and 6.8% higher than that of SignSpeaker and CG-Recognizer, respectively.

Key words: pre-training; gesture recognition; deep learning; electromyography(EMG); inertial measurement unit(IMU)

0　引　　言

手语是听障人士通过肢体动作和面部表情表达的主要交流形式,但难以让正常人所理解,这使得听障人士与正常人之间的交流存在障碍。手语识别技术有助于解决听障人士的沟通问题,因此近年来得到了广泛关注。

传统的手语识别方法通过人工提取特征并结

收稿日期:2023-04-07;**修回日期**:2023-05-24
基金项目:安徽省自然科学基金资助项目(2208085MF165)
作者简介:曹一丹(1998—),女,陕西咸阳人,合肥工业大学硕士生;
王青山(1975—),男,安徽合肥人,博士,合肥工业大学教授,博士生导师;
王　琦(1975—),女,安徽合肥人,博士,合肥工业大学副教授,硕士生导师,通信作者,E-mail:wangq@hfut.edu.cn.

合机器学习方法[1]来建立模型,采用隐马尔科夫、支持向量机、随机森林等模型。随着深度学习[2-3]在目标检测、图像分类等领域的突破性发展,越来越多的研究人员借助卷积神经网络(convolutional neural network,CNN)、循环神经网络(recurrent neural network,RNN)、长短期记忆(long short-term memory,LSTM)网络等来自动提取特征并识别。与传统机器学习方法相比,深度学习方法在大数据、大样本下处理效果更好,泛化能力更强。

基于传感器的手语识别系统中包括数据手套[4-5]、智能手表[6-7]、臂环[8-11]等设备。文献[4]开发了一种基于惯性测量单元(inertial measurement unit,IMU)的数据手套,使用陀螺仪和加速度计来获取方向、角度和加速度等运动数据,并识别 4 种方向手势;文献[5]开发了一种智能潜水手套,可以通过 5 个介质传感器捕捉水下手指的运动;文献[7]在三星 Galaxy Gear 智能手表上识别了 5 种手势,识别率为 87%,臂环结合了低成本的肌电信号(electromyography,EMG)和 IMU 的信号来跟踪手部运动和神经肌肉活动;文献[9]利用时域特征、均方根比和自回归模型提取表面肌电信号的特征,3 个通道的表面肌电信号可以准确地对 9 种手势进行分类。

传统特征提取方法泛化性较差。随着机器学习与深度学习的发展,越来越多的研究人员将之应用于数据的特征提取和分类。为了充分挖掘信号中的信息,文献[10]引入注意力机制[11]集成了不同尺度下的 EMG 信号特征。本文则是基于臂环进行研究。现有基于臂环的方法通常能获得很高的识别率,但局限于手势规模较小的情况,如 SignSpeaker[6]可以识别 103 种常用的美国手势,CG-Recognizer[8]利用频谱图构建特征生成器,对 50 种常用中国手语进行了实验,平均准确率超过 94%。随着手势规模的扩大,这些方法的识别率逐渐降低,主要原因是数据特征提取得不充分和泛化性差。

本文提出一个基于双路并行的大规模手势识别模型。设计一种基于 EMG 分支和 IMU 分支的双路并行模型来提取特征:EMG 分支设计双流结构,采用二维卷积神经网络构建空间、通道模块,重点关注 EMG 信号显著变化区域和重要通道;基于卷积长短时记忆(convolutional long short-term memory,ConvLSTM)网络[12]的 IMU 分支引入时间模块,专注于 IMU 信号中的时间信息;2 个分支在强调显著区域、突出重要通道、选择关键帧的同时融合 2 种信号特征。将预训练模型参数迁移至目标域,从而减少对目标域模型的训练,优化模型的训练效果,减少模型的训练成本。

1 双路并行手语识别模型

为了充分提取手势数据的特征,增强模型泛化性,从而提高手语手势识别率,本文提出一种基于 EMG 分支和 IMU 分支的双路并行手势预训练识别模型,如图 1 所示。

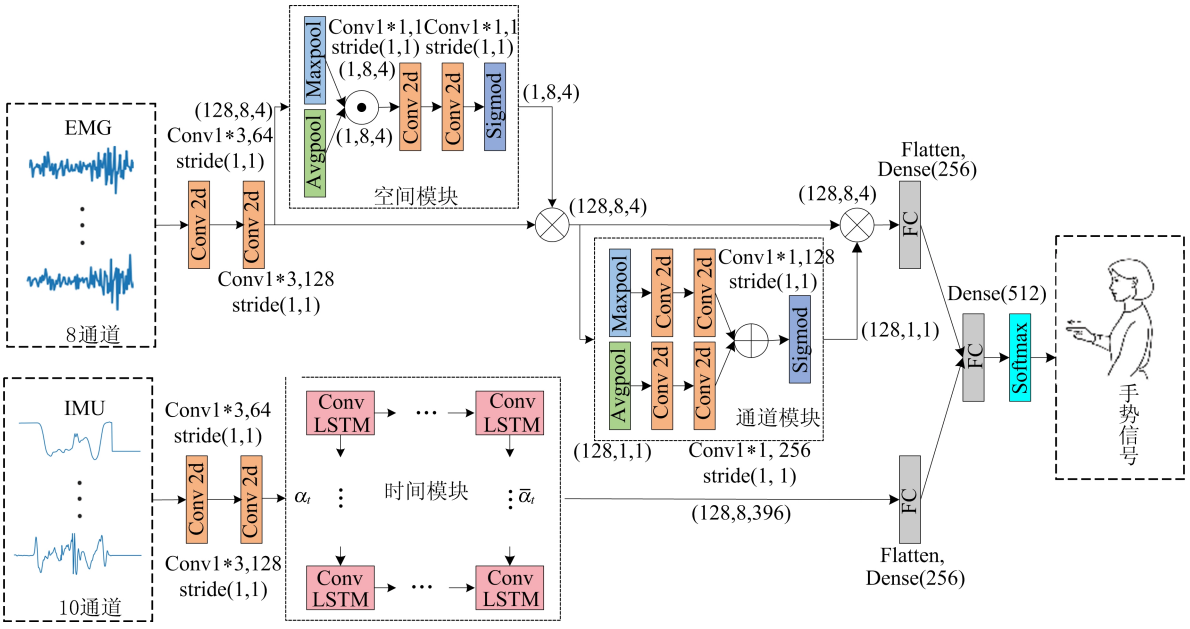

图 1 双路并行手势预训练识别模型

双路并行手势预训练识别步骤如下：① 对手语手势数据进行滤波并预处理为具有统一长度的数据作为输入；② 将数据输入到基于 EMG 分支和 IMU 分支的双路并行模型中，并进行模型的预训练；③ 利用训练好的模型参数进行参数迁移，迁移至目标域后进行微调。

在信号预处理方面：首先从数据中减去原信号的平均值以消除信号的偏移；再将信号输入到小波滤波器中以平滑信号；然后利用每个通道的平均能量作为阈值分割提取有用信号；最后将 EMG 和 IMU 提取的有用信号分别插值为具有统一长度的数据 $\boldsymbol{X}_{\mathrm{EMG}}$ 和数据 $\boldsymbol{X}_{\mathrm{IMU}}$。

1.1　EMG 分支

图 1 中，首先输入数据 $\boldsymbol{X}_{\mathrm{EMG}}$，经过两层卷积网络得到特征图 $\boldsymbol{F}_{\mathrm{EMG}}$；然后为了增强模型关注重点特征和重要通道的能力，本文进一步设计空间模块和通道模块。

1.1.1　空间模块

设计空间模块来强调 EMG 信号波动变化的显著区域并消除冗余信息。

1）将特征图分别通过 2 个分支变换到 2 个不同的特征空间 $\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Max}}$ 和 $\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Avg}}$ 上。上分支采取最大池化 Maxpool，下分支采取平均池化 Avgpool，即

$$\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Max}}=M(\boldsymbol{F}_{\mathrm{EMG}}) \tag{1}$$

$$\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Avg}}=A(\boldsymbol{F}_{\mathrm{EMG}}) \tag{2}$$

其中，M、A 分别表示最大池化操作和平均池化操作。2 种池化方法均能增强模型对特征的表示能力。

2）对 2 个分支 $\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Max}}$ 和 $\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Avg}}$ 按通道方向进行拼接，然后对池化特征进行卷积操作，将通道数量从 2 个减少到 1 个，即

$$\boldsymbol{C}=f(\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Max}}\odot\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Avg}}) \tag{3}$$

其中：$\odot$ 表示将池化结果按通道维拼接；f 表示卷积操作。

3）通过 Sigmod 激活函数生成空间图 $\boldsymbol{S}_{\mathrm{A}}$，即

$$\boldsymbol{S}_{\mathrm{A}}=\sigma(\boldsymbol{C}) \tag{4}$$

其中，σ 为 Sigmod 激活函数。

因此，经过空间模块作用后输出为：

$$\boldsymbol{F}_{\mathrm{S}}=\boldsymbol{F}_{\mathrm{EMG}}\boldsymbol{S}_{\mathrm{A}} \tag{5}$$

1.1.2　通道模块

设计通道模块是为了充分利用 EMG 信号的通道信息，注意显著的通道特征和抑制不必要的通道特征。

1）将 $\boldsymbol{F}_{\mathrm{S}}$ 分别输入到 2 个池化层和两层卷积神经网络中，再将得到的 $\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Max}}{}'$ 和 $\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Avg}}{}'$ 沿通道维进行相加，即

$$\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Max}}{}'=f(f(M(\boldsymbol{F}_{\mathrm{S}}))) \tag{6}$$

$$\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Avg}}{}'=f(f(A(\boldsymbol{F}_{\mathrm{S}}))) \tag{7}$$

$$\boldsymbol{C}'=\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Max}}{}'+\boldsymbol{F}_{\mathrm{EMG}}^{\mathrm{Avg}}{}' \tag{8}$$

2）在 Sigmod 函数后生成通道图 $\boldsymbol{C}_{\mathrm{A}}$，即

$$\boldsymbol{C}_{\mathrm{A}}=\sigma(\boldsymbol{C}') \tag{9}$$

3）将通道注意力的输出与输入相乘，得到新特征 $\boldsymbol{F}_{\mathrm{SC}}$，即

$$\boldsymbol{F}_{\mathrm{SC}}=\boldsymbol{F}_{\mathrm{S}}\boldsymbol{C}_{\mathrm{A}} \tag{10}$$

1.2　IMU 分支

手语手势数据属于时间序列数据，具有前后时间顺序信息，因此，IMU 分支在 ConvLSTM 基础上设计了时间模块，将 IMU 信号的空间信息与时间信息进行融合。

1）将输入数据 $\boldsymbol{X}_{\mathrm{IMU}}$ 经过两层卷积网络后得到特征图，即

$$\boldsymbol{F}_{\mathrm{IMU}}=f(f(\boldsymbol{X}_{\mathrm{IMU}})) \tag{11}$$

2）将 F_{IMU} 送入基于 ConvLSTM 网络的时间模块中，在保留空间联系的同时更加强调时序关系。时间权重设计如下：

$$\alpha_t=g(\boldsymbol{W}h_t) \tag{12}$$

其中：α_t 为第 t 时间步的时间权重；g 为激活函数；$\boldsymbol{W}$ 为 $1*1*1$ 的卷积核；h_t 为时间步 t 时的隐藏状态。然后，归一化权重可得：

$$\bar{\alpha}_t=\frac{\exp\alpha_t}{\sum\limits_t\exp\alpha_t} \tag{13}$$

EMG 分支和 IMU 分支经过全连接层输出相同维度的矩阵后再进行拼接，然后通过全连接层映射到手势规模维度，并使用 Softmax 函数映射到概率空间。

1.3　损失函数

双路并行手势识别模型的损失函数表示如下：

$$L=-\sum_{i=1}^{N}y_i\lg\hat{y}_i+\lambda_1\sum_j\|w_j^1\|^2+\lambda_2\sum_k\|w_k^2\|^2+\lambda_3\sum_t\|\bar{\alpha}_t\|^2 \tag{14}$$

其中：$\boldsymbol{y}=(y_1,y_2,\cdots,y_N)$，表示手语的真实类别；$\hat{\boldsymbol{y}}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_N)$，表示预测的向量；$N$ 为手语类别；w_j^1、w_k^2 分别为空间模块和通道模块的网络参数；λ_1、λ_2 为权值衰减参数；λ_3 为正则化系数，限制时间权重。

1.4　预训练模型

本文利用预训练[13]的思想，对部分志愿者手

势数据进行预训练,再将参数固定并迁移至新的志愿者手势数据中,以提高个体泛化性。预训练模型如图 2 所示。

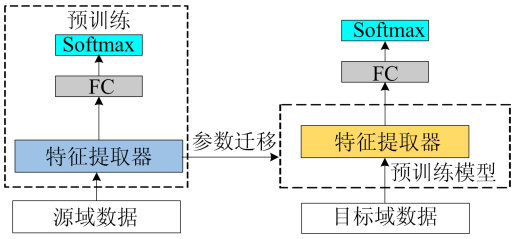

图 2 预训练模型

首先,按照志愿者将数据划分为源域数据集 D_S(7 名志愿者)和目标域数据集 D_T(4 名志愿者),在 D_S 上训练特征提取器,即为本文提出的双路并行手语识别模型;其次,固定特征提取器中训练好的参数,根据传递系数将训练好的参数传递给目标域网络,目标域网络的特征提取器仍为双路并行手语识别模型结构,目标域网络利用 D_T 微调参数以提高个体识别的泛化性。

2 实 验

本实验以右手为主利手,佩戴 MYO 手环采集手语手势的 IMU 信号和 EMG 信号。IMU 信号是一个 10 维信号,包括 4 维陀螺仪信号、3 维加速度信号和 3 维角速度信号,它能感知手臂的旋转、位移和其他位置信息;EMG 信号是 8 维向量,它记录了前臂 8 个部位的肌肉变化。IMU 信号的采样频率为 50 Hz;EMG 信号的采样频率为 200 Hz。

本文邀请了 4 名听障人士和 7 名听力健全人士,其中男性 6 名,女性 5 名,年龄在 18～45 岁之间。

根据《国家通用手语词典》[14],本文共采集了 500 种手语手势,每名志愿者每个手势做 20 次,共包含 110 000 个样本作为数据集。该数据集随机分为训练集(75%)和测试集(25%)。模型采用随机梯度下降的方式进行训练。学习率初始值设为0.001,逐渐衰减为 0.000 001。Batch size 设置为 8,采用 Adam 优化函数,使用 Dropout 方法防止模型过拟合。利用本文提出的双路并行手势识别模型对 EMG 分支和 IMU 分支分别进行训练。

对于 EMG 分支,损失函数中超参数权值衰减参数 λ_1、λ_2 满足 $\lambda_1>0$、$\lambda_2<1$ 并且 $\lambda_1+\lambda_2=1$。本文采用不同的 λ_1 进行实验,结果如图 3 所示。

从图 3 可以看出,当 $\lambda_1=0.7$ 时,EMG 分支的性能最佳。因此,在模型中设置 $\lambda_1=0.7$($\lambda_2=0.3$)。

对于 IMU 分支,本文用不同的正则化系数 λ_3 进行实验,结果见表 1 所列。从表 1 可以看出,$\lambda_3=0.010$ 时识别率最高。因此,在模型中设置 $\lambda_3=0.010$。

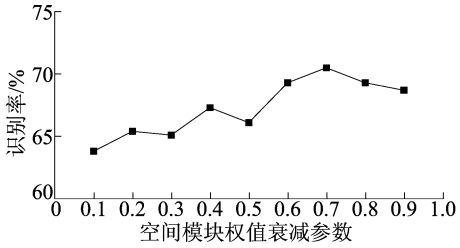

图 3 空间模块权值衰减参数设置实验结果

表 1 正则化系数设置实验结果

λ_3	识别率/%
0.001	70.3
0.005	60.4
0.010	75.1
0.050	65.9
0.100	67.7
0.500	62.4

2.1 消融实验

为了验证模型中每个模块的作用,本节进行消融实验。消融实验的对照组设计如下:空间模块,通道模块,空间模块+通道模块,空间模块+时间模块,通道模块+时间模块,空间模块+通道模块+时间模块。

消融实验结果如图 4 所示。从图 4 可以看出,3 种模块均同时包含时效果最好,其原因是在可以充分提取信号特征的同时,从 3 种维度强调重要特征。

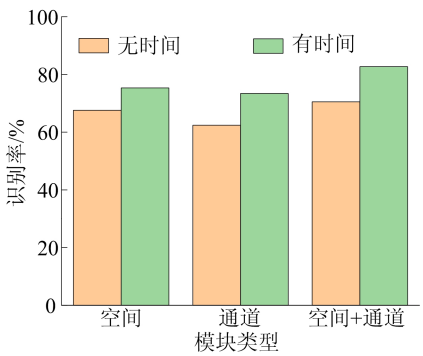

图 4 消融实验结果

2.2　比较实验

2.2.1　与基线比较

进行基线比较实验的方法包括传统机器学习方法[1]和深度学习方法[2-3]，由隐马尔科夫、支持向量机、随机森林、CNN、RNN、LSTM 模型组成。基线比较实验结果见表 2 所列。

从表 2 可以看出，本文模型的识别率显著高于其他模型。这是由于本文模型同时考虑到手势信号的空间、通道、时间信息，从而提高了模型的识别率。

表 2　基线比较实验结果

模型	识别率/%
隐马尔科夫	15.6
支持向量机	16.3
随机森林	19.1
CNN	23.7
RNN	20.5
LSTM	30.2
本文模型	82.1

2.2.2　与现有技术比较

将本文模型与 SignSpeaker[12]、CG-Recognizer[4]进行比较，实验结果见表 3 所列。

从表 3 可以看出，本文模型的性能上优于其他 2 种模型。与 SignSpeaker、CG-Recognizer 相比，本文模型将平均识别率分别提高了 21.0%、6.8%。这是由于本文提出的模型通过双路分支提取信号的空间、通道和时间信息来全面刻画手语手势特征，并且利用预训练模型提高模型的泛化性。

表 3　现有技术比较实验结果

模型	识别率/%
Signspeaker	61.1
CG-Recognizer	75.3
本文模型	82.1

此外，为验证模型稳定性，分别从 500 个手势中任意选择 10、50、100、200、300、400 个手势进行实验，结果如图 5 所示。

从图 5 可以看出：本文模型和 CG-Recognizer 都具有较好的识别性能，在手势规模较小时，本文模型的识别率略高于 CG-Recognizer 和 LSTM；当手势规模增加时，CG-Recognizer 和 LSTM 的识别性能下降，而本文模型的识别性能保持较高水平。实验结果表明，本文模型在大规模手势识别中是有效的。

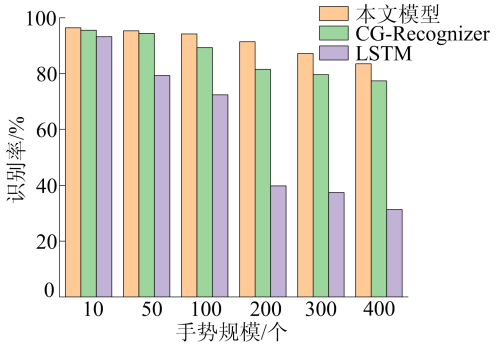

图 5　稳定性实验结果

2.3　泛化性实验

为了评价本文模型的个体泛化性，本文选择 4 名新的志愿者对模型进行评估，泛化性实验结果见表 4 所列。每位志愿者对 500 个手势每个做 20 遍，共收集 40 000 个样本。

表 4　泛化性实验结果

志愿者	识别率/%
志愿者 1	80.2
志愿者 2	78.6
志愿者 3	78.0
志愿者 4	79.5
平均识别率	79.1

表 4 显示了 4 名志愿者在本文模型上的平均识别率，结果表明该模型具有用户泛化能力，能够识别新用户的手语手势。

3　结　　论

本文提出了一种基于双路并行的大规模手势识别模型。EMG 分支设计双流结构，采用二维卷积神经网络构建空间、通道模块，重点关注 EMG 信号显著变化区域和重要通道；IMU 分支在 ConvLSTM 基础上引入了时间模块，将时间信息融于空间信息之中。本文模型能充分提取数据空间、通道、时间特征，并通过模型预训练增强了模型的泛化能力。

[参　考　文　献]

[1]　GALEA L, SMEATON A F. Recognising irish sign language using electromyography [C]//2019 International Conference on Content-Based Multimedia Indexing. [S. l.: s. n.], 2019: 1-4.

（下转第 605 页）

ment elastic-plastic model for subsurface initiated spalling in rolling contacts[J]. Journal of Tribology, 2014, 136(1): 1-14.

[19] WEN J S, JU W E, HAN T K, et al. Finite element analysis of a subsurface penny-shaped crack with crack-face contact and friction under a moving compressive load[J]. Journal of Mechanical Science and Technology, 2012, 26(9): 2719-2726.

[20] 党晓勇，师志峰，刘静. 圆柱滚子轴承次表面裂纹区域应力分布规律研究[J]. 机电工程，2023，40(2)：204-210，224.

[21] BELYTSCHKO T, BLACK T. Elastic crack growth in finite elements with minimal remeshing[J]. International Journal for Numerical Methods in Engineering, 1999, 45: 602-620.

[22] HANSBO A, HANSBO P. A finite element method for the simulation of strong and weak discontinuities in solid mechanics[J]. Computer Methods in Applied Mechanics & Engineering, 2004, 193(33/35): 3523-3540.

[23] SONG J, AREIAS P M A, BEIYTSCHKO T. A method for dynamic crack and shear band propagation with phantom nodes[J]. International Journal for Numerical Methods in Engineering, 2006, 67(6): 868-893.

[24] 伍梓聪，任祉达，李蓓智. 滚动轴承微裂纹及其对轴承疲劳寿命的影响[J]. 组合机床与自动化加工技术，2020，8(8)：21-24，29.

（责任编辑　胡亚敏）

（上接第 589 页）

[2] BIRD J J, KOBYLARZ J, FARIA D R, et al. Cross-domain MLP and CNN transfer learning for biological signal processing: EEG and EMG[J]. IEEE Access, 2020, 8: 54789-54801.

[3] FUKANO K, IIAZAWA K, SOEDA T, et al. Deep learning for gesture recognition based on surface EMG data[C]// 2021 International Conference on Advanced Mechatronic Systems. [S. l. : s. n.], 2021: 41-45.

[4] MAKAUSSOV O, KRASSAVIN M, ZHABINETS M, et al. A low-cost, IMU-based real-time on device gesture recognition glove[C]//2020 IEEE International Conference on Systems, Man, and Cybernetics. [S. l.]: IEEE, 2020: 3346-3351.

[5] ANTILLON D W O, WALKER C R, ROSSET S, et al. Glove-based hand gesture recognition for diver communication[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(12): 1-13.

[6] HOU J, LI X, ZHU P, et al. SignSpeaker: a real-time, high-precision smartwatch-based sign language translator[C]// The 25th Annual International Conference on Mobile Computing and Networking. [S. l. : s. n.], 2019: 1-15.

[7] WEN H, RAMOS ROJAS J, DEY A K. Serendipity: finger gesture recognition using an off-the-shelf smartwatch[C]// Proceedings of CHI Conference on Human Factors in Computing Systems. [S. l. : s. n.], 2016: 3847-3851.

[8] ZHENG Z, WANG Q, YANG D, et al. CG-Recognizer: a biosignal-based continuous gesture recognition system[J]. Biomedical Signal Processing and Control, 2022, 21(7): 2398-2410.

[9] DUAN F, REN X, YANG Y. A gesture recognition system based on time domain features and linear discriminant analysis[J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 13(1): 200-208.

[10] SUN B, SONG B, LV J. A multi-scale feature extraction network based on channelspatial attention for electromyographic signal classification[J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 15(2): 591-601.

[11] PAN T Y, TSAI W L, CHANG C Y. A hierarchical hand gesture recognition framework for sports referee training-based EMG and accelerometer sensors[J]. IEEE Transactions on Cybernetics, 2022, 52(5): 3172-3183.

[12] HUANG H, ZENG Z, YAO D, et al. Spatial-temporal ConvLSTM for vehicle driving intention prediction[J]. Tsinghua Science and Technology, 2022, 27(3): 599-609.

[13] 李晓林，胡泽荣. 基于预训练模型的中文电子病历实体识别[J]. 计算机工程与设计，2023，44(2)：535-540.

[14] 中国聋人协会，国家手语和盲文研究中心. 国家通用手语词典[M]. 北京：华夏出版社，2019：5-231.

（责任编辑　胡亚敏）