DOI:10.19651/j.cnki.emt.2312748

基于对比学习的信息缺失手势识别新方法*

卞雨玮[1,2] 华立涛[1,2] 周 媛[1,2]
(1.南京信息工程大学人工智能学院 南京 210044; 2.南京信息工程大学未来技术学院 南京 210044)

摘 要:针对现有深度学习模型识别信息缺失手势需要大量标注数据、更深的网络需要更多参数的问题,首先收集整理了一个信息缺失手势数据集 IMG_NUIST,然后借鉴对比学习思想,提出了一个新的信息缺失手势识别模型 CLGR,该模型通过对手势类内和类间差异度的对比约束提高模型特征学习性能。在两个经典数据集(ASL Alphabet 和 NUS I)和新提出的 IMG_NUIST 数据集上进行了广泛实验,消融实验表明对比学习思想能有效地将平均识别准确率提高至 98.60%以上且收敛速度显著提升;对比实验表明本文所提模型计算复杂度比其它 4 个模型平均简化了 41.4%,在 NUS I 和 IMG_NUSIT 数据集上的手势识别准确率超过四个对比方法,特别是在 NUS I 数据集上将识别准确率平均提高了 17.35%,在 ASL Alphabet 数据集上的识别准确度仅比最优结果低 0.43%。实验结果说明所提模型对于缺失手部部分信息和杂乱背景等问题的手势识别任务有显著效果,具有收敛速度更快、计算复杂度更少的优秀性能,有很好的实用价值。

关键词:信息缺失;杂乱背景;手势识别;深度学习;对比学习
中图分类号:TP3 **文献标识码**:A **国家标准学科分类代码**:520.2

Novel method for gesture recognition with missing information based on contrastive learning

Bian Yuwei[1,2] Hua Litao[1,2] Zhou Yuan[1,2]
(1. School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China;
2. School of Future Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: Aiming at the problem that the information missing gestures recognition based on deep learning needs a large amount of labeled. The deeper the network needs more parameters, we first collect a data set called IMG_NUIST which consists of information missing gestures and full gestures. Then we propose a new gesture recognition model CLGR, the inter-class and intra-class similarities constraints enhance the feature learning performance of the model. Extensive experiments are conducted on two classic datasets (ASL Alphabet and NUS I) and the proposed IMG_NUIST dataset. The experimrnt results are shown as follows: 1) in the ablation study, contrastive learning can effectively improve recognition accuracy up to over 98.60% and the model convergence speed are significantly accelerated. 2) In the comparative experiments with two recent works and two contrastive learning models, the computational complexity of CLGR is 41.4% simpler than that of the two comparison models on average. CLCR can recognize the gestures with missing information and works well for those gestures with cluttered backgrounds. The gesture recognition accuracy of CLCR on the NUS I and IMG_NUSIT data sets outperforms the four comparison methods and is only 0.43% lower than the best result on ASL. Especially on the NUS I dataset, CLCR increases the recognition accuracy of gestures by 17.35% on average. The experimental results show that the proposed model is significantly effective for gesture recognition tasks with missing information and cluttered background with fast convergence speed and low computational complexity, and it is practical.

Keywords: missing information; cluttered backgrounds; gesture recognition; deep learning; contrastive learning

0 引 言

手势是人机交互的一种方式,在人与机器之间搭建一座桥梁。手势识别任务有着广泛的应用[1-2],如帮助聋人进行日常交流、辅助驾驶系统等,吸引了众多研究者的关注。手势识别方法可分为两类:基于传感器数据的方法[3-6]和基

收稿日期:2023-02-06
* 基金项目:江苏省大学生创新训练一般项目(202110300094Y)资助

于视觉的方法[7-9]。基于传感器的手势识别是利用数据手套或鼠标来获取或跟踪用户的手部信息以进行识别,该方法具有输入数据少、速度快的优点,但可穿戴设备牺牲了用户的便利性,并且昂贵的设备增加了识别成本。相比之下,基于视觉的识别为用户提供了更大的自由度和灵活性,它通过深度神经网络对手势图像进行识别[10-12],降低了识别维度并提高了识别的灵活性,目前基于视觉的深度神经网络方法成为手势识别的主流。

然而,基于深度学习的手势识别仍然面临着许多挑战:现实中常见手与相机垂直或偏移一定角度时,手势会在一定程度上不完整,影响识别精度;杂乱的背景会干扰手势识别的效果。目前针对这些非理想手势识别鲜见报道。进一步的,现有公开手势数据集中都是完整的手势图片,缺少信息不完整的手势数据集;并且深层的网络模型参数量和计算量过大,不利于模型向边缘端迁移、收敛速度慢,这些都影响了手势识别的实际落地应用。因此,寻求一种轻量级模型有效识别非理想情况的手势仍然是值得研究的课题。

现有公开的手势数据集中只涉及完整手势数据并不包含信息缺失的手势数据,而在实际生活中,获取的手势图像往往达不到公开数据集的标准,如手势不全、背景干扰等。因此,我们采集并整理了缺失信息手势数据,并构建了一个新的手势数据集南京信息工程大学信息缺失的手势数据集(information missing gestures of NUIST,IMG_NUIST),完善了现有的手势数据集并为相关研究提供测试基准。

受对比学习思想的启发,本文提出一种新的信息缺失手势识别模型,主要贡献可以归纳为以下4点:

1)提出了一种基于对比学习思想的手势识别新方法,这种轻量化的模型能够鲁棒地在非理想条件下准确高效地识别手势;

2)提出了一种增强的相似性损失 $\mathcal{L}_{sim}$ 来度量样本之间的距离,提高了训练稳定性和收敛速度;

3)收集并整理了包含缺失信息手势的新数据集IMG_NUIST,以补充现有手势数据;

4)进行了广泛大量的实验,分析并验证了所提方法的性能。

1 相关工作

1.1 深度神经网络

深度神经网络因其端到端的学习方式打破了传统机器学习手工提取图像特征的局限性,已被广泛应用于手势识别任务[13-17]。

2019年,Cheng等[18]提出了一种基于卷积神经网络和受限玻尔兹曼机的联合网络架构,该模型通过Kinect传感器获取10类手势样本,将多重受限玻尔兹曼机无监督提取的手势特征和卷积神经网络有监督提取的手势特征相结合,模型对这10类手势的识别错误率仅为3.9%。同年萧炎武提出了一种双通道网络卷积神经网络(DC-CNN)[19],DC-CNN利用Canny边缘检测技术对原始图像进行手部边缘检测,接着将手势图像和手部边缘图像作为模型两个通道的输入,每个通道都有一个单独的权重并对应一个softmax分类器。程换新等[7]结合残差架构和图卷积模型构建双流网络,将手势构建成空间图和时序图作为模型输入,在CSL和DEVISIGN-L手势数据集上识别精度分别达到了96.2%和69.3%。2020年,Sun等[20]提出了一种基于双流卷积神经网络多级特征融合模型,首先对手势进行目标检测再识别手势,该模型在光照变化、背景杂乱和部分遮挡场景下对手势的平均检测精度提高了1.08%,手势识别平均精度提高了3.56%。

最新的手势识别工作是2021年Tan等[21-22]提出的EDenseNet[21]和CNN-SPP[22],这两个工作都是基于卷积神经网络(convolutional neural networks, CNN)。EDenseNet是一种增强型密集连接卷积神经网络,网络中某层特征是通过将前面多层特征短接后和当前层特征拼接得到,随后连接的卷积层平滑掉不需要的特征得到最终特征,根据此特征进行手势识别;CNN-SPP是一种集成空间金字塔池化(SPP)的卷积神经网络(CNN),用堆叠池化代替单个池化对,对学习到的特征进行堆叠池化后输入全连接层,最后全连接层输出的特征被用于手势识别。

1.2 对比学习

对比学习[23-24]是一种自监督学习思想,其核心是通过最小化相同实例距离、最大化不同实例差异学习一个特征编码器,通过使不同类别数据的编码结果尽可能不同达到增强特征编码器能力。由于对比学习是在抽象语义层面区分样本无需关注实例繁琐细节,因此对比学习模型更简单,泛化能力更强。

2020年,He等[25]提出了一种无监督视觉表示模型——动量对比(momentum contrast,MoCo),MoCo框架主要有两个编码器(键值编码器和查询编码器),均采用初始化完全相同的ResNet作为主干网络。“教师—学生”模式下,查询编码器采用常规的梯度下降法更新参数,“学生”键值编码器从“教师”查询编码器”获得最新参数并采用动量更新法进行缓慢平滑地更新自己参数。其创新点之一是引入了动态字典来存储编码器对各样本的编码结果,用训练样本编码结果和动态字典中键值的相似性来计算对比损失以训练查询编码器。MoCo的两个核心思想是:1)将字典视为动态队列,字典更新大小与模型输入样本批量大小相同,每一批新的编码进入字典队列时,最早的相同大小批量的编码退出队列;2)MoCo以动量方式更新字典编码器,使得参数缓慢平滑地被更新,解决了由于字典包含很多样本,每次反向传播时字典编码器变化太大,导致提取的特征一致性低。基于“同目标不同视角的输入,其视觉表示应该具有不变性”的事实,Chen等[26]在2020年提出了视觉表征的对比学习(SimCLR)框架。simCLR对输入图像进行两次随机数据增强来模拟来自同一图像的两个不同视角的输

入，通过编码器对这两幅图像进行特征提取并线性投影，使用余弦相似度来度量特征投影的相似度。通过最大化同一目标不同视角的相似性、最小化不同目标之间的相似性来训练模型，但是 SimCLR 需要大量的计算能力。随后，何凯明等在 SimCLR 的基础上提出了 MoCo-v2[27]，通过在 MoCo 中加入特征投影和更多数据增强，建立了一个比 SimCLR 性能更好、更强大的模型，并且不需要大规模的批量训练，极大地解放了计算能力。

2　本文方法

同类缺失手势和完整手势虽然视觉上信息量不同，但是能够在语意层面完全一致，所以可以通过对比学习来拉近同类缺失手势和完整手势的特征编码距离。但是对比学习模型只关注个体自身特征不考虑个体所属类别的特征，直接用于手势识别任务会导致性能不佳。因此，本文受 MoCo 启发，将对比学习思想引入信息缺失手势识别任务，提出一个基于对比学习的信息缺失手势识别新方法(contrastive learning based gesture recognition，CLGR)，该模型通过对完整手势数据进行处理得到对应的缺失手势图像，通过缺失手势和完整手势之间的对比学习，在编码器编码过程中能够在语意层面拉近缺失手势和完整手势距离、补充缺失手势信息，以增强特征编码器对缺失手势的表示能力。

图 1 是本文模型 CLGR 架构，主要包含 3 个部分：手势特征编码器、手势分类器和相似度比较器。首先手势特征编码器对随机采样的完整手势和信息缺失手势分别进行编码，然后将完整手势特征更新相似度比较器中的动态字典，手势分类器对信息缺失手势特征进行分类，同时相似度比较器计算信息缺失手势特征和动态字典中所有完整手势特征的相似度损失，最后分类损失和相似度损失共同被用于优化手势特征编码器。

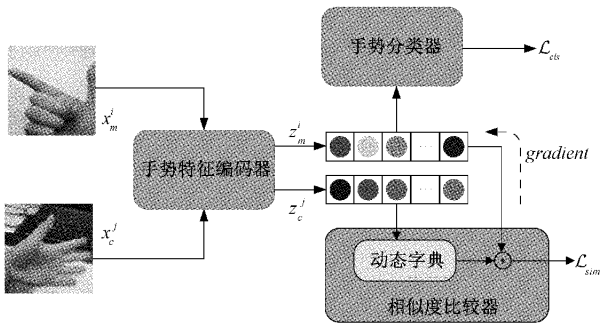


图 1　基于对比学习的信息缺失手势识别模型

CLGR 主要有 4 处和 MoCo 不同：1)输入数据要求不同，MoCo 要求进入不同编码器的图像出自同一个对象(同一类)，而所提 CLGR 是对信息缺失手势和完整手势进行随机采样获取模型输入，不要求输入类别对应，约束更松弛；2)编码器数量减少，MoCo 采用键值编码器和查询编码器双编码器架构，而 CLGR 中仅采用一个手势特征编码器，模型复杂度降低；3)动态字典大小设置不同，MoCo 的动态字典大小为一个固定超参(默认值为 65 536)，每次更新量和训练样本批量大小一致，而 CLCR 中动态字典大小和任务待识别种类数目线性相关，存储空间需求大大降低；4)将同一类别的类内相似度引入了代价函数。

1)手势特征编码器

CLCR 的输入有两个：缺失手势 $\boldsymbol{x}_m$ 和完整手势 $\boldsymbol{x}_c$，x_m^i 和 x_c^i 分别是对缺失手势和完整手势的随机采样。CLCR 以 ResNet18 为编码器主干网，连接一个 128 维的全连接层作为输出层，手势特征编码器对输入图像进行编码，得到手势特征 $z = f(\boldsymbol{x}) \in \mathbb{R}^d$，$f(\cdot)$表示通过编码器对手势进行特征编码，则两个输入通过手势特征编码器分别输出特征向量 z_m^i 和 z_c^i。

2)手势分类器

本文采用一个全连接层跟上一个 softmax 作为手势分类器，对编码器输出的缺失手势特征进行分类。对应的损失函数如式(1)所示。

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N} p(c_l \mid \boldsymbol{z}_m^i) \tag{1}$$

其中，$\boldsymbol{x}_m^i$ 表示第 i 个缺失手势样本，$p(k \mid x_m^i)$ 表示样本 x_m^i 属于第 c_l 类的概率，N 是批量大小。

3)相似度比较器

相似度比较器的核心是动态字典，动态字典被用来存储动态更新的完整手势特征。为保证字典中数据分布平衡，在动态字典中每类手势都存储相同数量的特征。假设当前信息缺失手势 $\boldsymbol{x}_m^i$ 所属类别为 c_l，定义类内相似度 sim_{inter}^i 和类间相似度 sim_{intra}^i。sim_{inter}^i 是度量缺失手势 x_m^i 和动态字典中同类完整手势的特征相似度，sim_{intra}^i 是度量缺失手势 x_m^i 和动态字典中所有不同类的完整手势的特征相似度：

$$sim_{inter}^i = \frac{1}{|c_l|}\sum_{j=1}^{|c_l|} \mathrm{e}^{\alpha < f(\boldsymbol{x}_m^i), f(\boldsymbol{x}_{c_l}^j) >} \tag{2}$$

$$sim_{intra}^i = \sum_{k \neq l}^{|c|} \frac{1}{|c_k|}\sum_{j=1}^{|c_k|} \mathrm{e}^{\alpha < f(\boldsymbol{x}_m^i), f(\boldsymbol{x}_{c_k}^j) >} \tag{3}$$

其中，x_m^i 属于第 c_l 类，$x_{c_l}^j$ 和 $x_{c_k}^j$ 分别表示第 c_l 类的第 j 个完整手势和第 c_k 类的第 j 个完整手势，$|\ c\ |$ 表示手势类别总数，α 为缩放因子，以此保证通过指数匀运算得到的相似度保持在合理范围，设置为 0.01，$|\cdot|$ 表示该类样本数据量，$<\cdot>$表示内积运算。

期望相同类别的特征趋于一致，不同类别的特征相似度趋于 0。因为在特征层采用 sigmoid 函数作为激活函数使得特征编码每一维范围是 0～1，手势特征两两进行内积运算的结果上界等于特征向量的长度 d，在缩放因子的作用下，有：

$$0 \leqslant e^{\alpha<f(x_m^i),f(x_{e_l}^j)>} \leqslant e^{ad} \tag{4}$$

$$0 \leqslant e^{\alpha<f(x_m^i),f(x_{e_k}^j)>} \leqslant e^{ad} \tag{5}$$

因此考虑类内和类间样本的特征分布,最终定义一个新的相似度损失 $\mathcal{L}_{sim}$:

$$\mathcal{L}_{sim} = -\log \frac{1}{N}\sum_{i=1}^{N} \frac{sim_{inter}^{i}}{sim_{intra}^{i} + sim_{inter}^{i}} \tag{6}$$

从式(4)和(5),有:

$$0 < \frac{sim_{inter}^{i}}{sim_{intra}^{i} + sim_{inter}^{i}} \leqslant 1 \tag{7}$$

由式(7)可知,通过最小化类内相似度、最大化类间相似度可以实现模型训练,所以相似度损失的实质就是最小化 $\mathcal{L}_{sim}$。

4)模型损失函数

最终模型总的损失函数是联合相似度损失和分类损失:

$$L = L_{sim} + L_{cls} \tag{8}$$

3 实验及分析

此外,我们还在两个经典数据集美国手语字母数据集(American Sign Language Alphabets set, ASL Alphabet)[28]和新加坡国立大学手势数据集 I(National University of Singapore,NUS I)[29]进行了实验,其中 ASL Alphabet 数据集包含了 26 个英文字母、空格、删除等共 29 类 87 000 张图片,NUS I 数据集包含了 10 类手势共 2 000 幅图像。为了验证模型对缺失手势的识别效果,沿着手势中心对两个数据集图像进行裁剪,并人工剔除没有任何手势信息的图像,最后得到这个两数据集对应的缺失手势数据。图 2~4 分别是 IMG_NUIST 数据集、ASL Alphabet 数据集和 NUS I 数据集的部分样本,第一行是完整手势,第二行是缺失手势。特别值得注意的是,NUS I 数据集背景干扰最大。

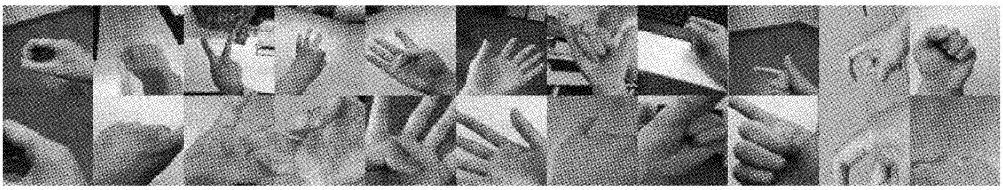
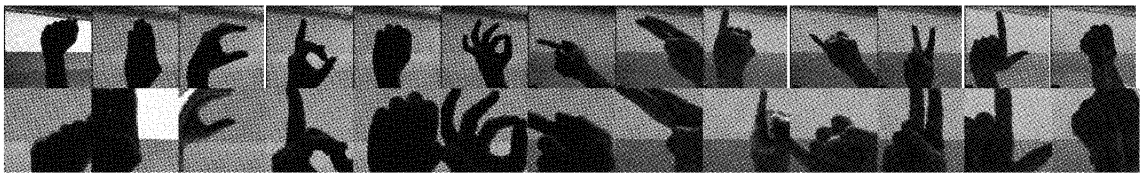
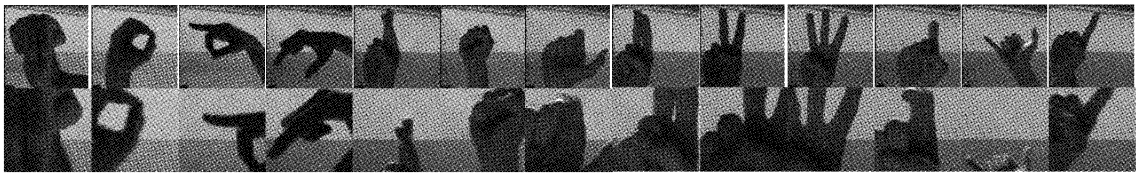


图 2 IMG_NUIST 数据集



(a) a~m的手势表示(从左到右)



(b) n~z的手势表示(从左到右)

图 3 ASL Alphabet 数据和对应缺失手势数据

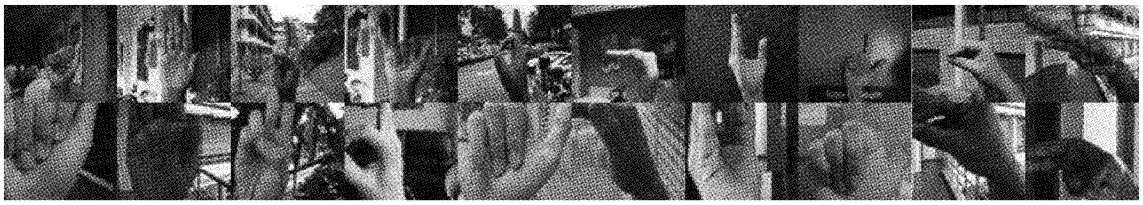


图 4 NUS I 数据和对应缺失手势数据

本文通过缺失手势特征和动态字典中的完整手势特征比较获得相似度损失,并结合分类损失得到综合损失,以最小化综合损失作为目标来指导特征编码器的训练。

3.1 实验细节

为了确保训练数据中各手势数据分布均匀,通过数据增强确保每个数据集中不同类别的手势数据都有相同数量,手势特征编码器输出特征 z 大小为 128,动态字典中每

类手势特征个数都设定为 64，对 11 类手势识别任务来说，动态字典大小即为 64 × 11 = 704。实验在 NVIDIA GTX1070 上进行，采用 Python3.7.6 加 Tensorflow2.6.0 框架。训练时 bathcsize 设置为 64，学习率设置为 0.001，训练轮次设置为 100。

值得注意的是，本文 CLCR 模型的输入是信息缺失手势和完整手势，每次迭代时交替从信息缺失手势数据和完整手势数据中随机采样传入模型。

3.2　消融实验

为了验证所提出的相似性损失 $\mathcal{L}_{sim}$ 的有效性，我们进行了 5 折交叉验证的消融实验，结果如表 1 所示。

表 1　三个数据集上的五折交叉验消融实验结果

数据集	实验轮次	准确率/%		收敛速度(轮次)	
		无 $\mathcal{L}_{sim}$	有 $\mathcal{L}_{sim}$	无 $\mathcal{L}_{sim}$	有 $\mathcal{L}_{sim}$
IMG_NUIST	1	99.91	99.98	9	7
	2	99.85	100	8	8
	3	99.88	100	10	7
	4	99.81	100	9	6
	5	99.93	100	8	9
	平均	99.87	100	8.8	7.4
ASL Alphabet	1	96.59	99.23	35	14
	2	95.95	99.65	28	20
	3	94.71	99.26	34	21
	4	94.81	99.43	27	12
	5	94.52	99.55	25	19
	平均	95.32	99.42	29.8	17.2
NUS I	1	97.33	97.33	15	9
	2	96.83	98.83	18	8
	3	93.17	99.00	16	11
	4	95.67	98.67	20	9
	5	97.00	99.17	18	10
	平均	95.60	98.60	17.4	9.4

加入相似度损失的 CLGR 在 3 个数据集上的收敛速度明显高于未加入相似度损失的基准模型 ResNet18[30]，并且对手势的识别率也得到了提高。特别是对更复杂的 ASL Alphabet 数据集和 NUS I 两个数据集，平均训练收敛速度分别提高了 42.3%和 46.0%，对缺失手势的平均识别准确率分别提高了 4.1%和 3.0%。值得注意的是，虽然 CLGR 在 IMG_NUIST 数据集上的平均训练收敛速度和平均识别准确率和基准相差不大，但是训练过程却更加稳定。图 5 显示了训练收敛过程，图 5(a)、(b)、(c)分别表示在 IMG_NUIST，ASL Alphabet，NUS I 三个数据集上不同算法的收敛过程。可见本文模型在三个数据集上更鲁棒、更稳定，尤其是在 IMG_NUSIT 上，基准模型震荡激烈而 CLGR 表现出了优越的稳定性。

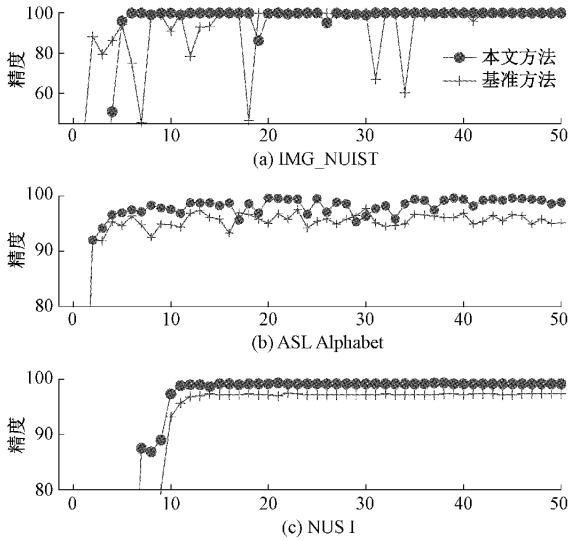


图 5　不同数据集上训练收敛曲线

为利用完整手势数据实现杂乱背景及信息缺失的手势识别任务，首先我们采集了各种背景下(白天、夜晚、白墙、办公室等)数字 0～10 的完整手势数据，然后对这些数据进行随机裁剪获得相应的缺失手势。最后整理出包含 21 614 张图像的 IMG_NUIST 数据集①(如图 2 所示)，数据集中信息缺失的手势数据可为识别不完整手势研究提供验证，该数据集补充了现有手势数据集并为传统手势识别任务提供新数据。

3.3　对比实验

为了全面验证所提方法的性能，我们将 CLGR 与两个最新手势识别工作 EDenseNet[21] 和 CNN-SPP[22]，以及两个对比学习模型 simCLR[26] 和 MoCo v2[27] 在三个数据集上进行了手势识别性能比较(如表 2 所示)。采用参数量、浮点计算数 FLOPs(floating point operations)和平均准确率来定量评价模型大小、模型复杂度和识别性能，其中参数量和 FLOPs 数值越小越好。

表 2 中，CLGR 模型的参数量远低于两个对比学习模型 simCLR 和 MoCo v2，计算复杂度 FLOPs 相比减少了 20%，并且在 3 个数据集上的识别率平均提高 32.40%。直接用自监督对比学习算法 simCLR 和 MoCo v2 识别信息缺失手势和完整手势的最高准确率不超过 80%，本文提出的方法借助对比学习思想，通过引入类内差异最小化“迫使”模型提取信息缺失手势的有效类别特征，从而提升分类性能。

和基于 CNN 的模型相比，CLGR 的参数量比 CNN-SPP 减少了 23.5%，模型复杂度 FLOPs 比 EDenseNet 减少 42.2%、比 CNN-SPP 减少了 83.0%；对比这两个模型，

① IMG_NUIST 数据集网盘地址：https://pan.baidu.com/s/19TGTxQ3RDmJ5wlUK55n-KA? pwd=1234，提取码：1234

表 2 三种方法在三个数据集上结果(非理想手势指信息缺失手势和杂乱背景下的手势)

方法	参数量/M	FLOPs/G	IMG_NUIST		ASL Alphabet		NUS I	
			非理想手势识别准确率/%	完整手势识别准确率/%	非理想手势识别准确率/%	完整手势识别准确率/%	非理想手势识别准确率/%	完整手势识别准确率/%
EDenseNet	1.09	2.51	99.83	99.40	**99.85**	**99.60**	96.88	92.50
CNN-SPP	6.37	8.52	99.35	99.77	99.76	99.42	95.15	94.50
simCLR	11.5	1.82	69.14	71.05	68.95	68.25	61.22	69.44
MoCo v2	11.69	1.82	64.26	64.45	60.80	64.49	71.88	68.75
CLGR	4.87	**1.45**	**100**	**99.99**	99.42	99.04	**98.60**	**98.67**

CLGR 的参数量虽然不是最少的,但其浮点计算量却是最少的,非理想环境下,CLGR 在 IMG_NUIST 和 NUS I 这两个数据集上取得了最高手势识别准确率,尤其是在杂乱背景的 NUS I 上,EDenseNet 和 CNN-SPP 的准确率不到 97%,而 CLGR 识别准确率却高达 98.6%。在识别完整手势时,CLGR 在训练过程中仅用信息缺失手势的标签与预测值计算交叉熵损失,使得模型在 IMG_NUIST 上预测完整手势时准确率有所减少,但依旧是取得了最高的准确率,而在 NUS I 数据集上预测完整手势时,由于该数据集中完整手势包含杂乱背景信息相较于信息缺失手势更多,使得 EDenseNet 和 CNN-SPP 的准确率都有所下降,尤其 EDenseNet 的准确率下降了 4.38%,而 CLGR 依旧保持稳定的识别准确率并位列第一。

这边本文模型在 ASL Alphabet 上的识别率不是最高,主要原因是在 ASL Alphabet 数据集上部分手势的视觉差距较小,如图 6 中的“i”和“j”对应的手势,所以 CLCR 学到的“i”和“j”的视觉特征类似。但 CLGR 在最低模型复杂度情况下获得了和另外两个工作相近的识别性能,仅比最优的 EDenseNet 低了 0.43%。

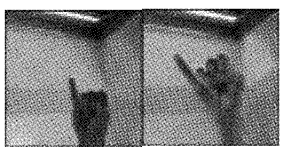


图 6 ASL Alphabet 数据集中“i”和“j”的手势

因为 CLGR 和 simCLR、MoCo v2 都是用 ResNet18 作为骨干网,所以它们 3 个推理时间一致,因此只对比了 CLGR、EDenseNet 和 CNN-SPP 的推理性能。图 7 展示了 CLGR 与这两种模型在不同批量大小设置下推理时间的对比图,可以看出当批量大小设置较小时,CLGR 与其余两种算法的推理时间相差较小,但随着批量大小的增加,到每个批量达到 64 时,CLGR 在推理速度上展现出极佳的优势。并且模型的推理损耗时间与模型浮点计算量成正比,而 CLGR 是 3 个模型中 FLOPs 最小的,因此相较于其它两个模型,CLGR 的推理速度最快,该优势随着预测数据量的增加变得尤为明显。

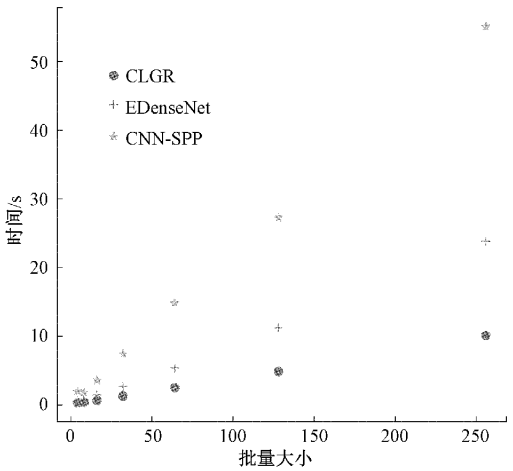


图 7 三种方法在不同批量大小下的推理时间比较

4 结 论

本文针对信息缺失手势识别任务,采集整理了一个新的数据集 IMG_NUIST,并提出了一种基于对比学习策略的新方法 CLGR。该方法通过缺失手势特征和动态字典中的完整手势特征比较获得相似度损失,并结合分类损失得到综合损失来指导特征编码器的训练,提高了手势识别的稳定性和速度。在 3 个数据集上进行了广泛的实验,消融研究表明所提相似性损失的有效性,对比实验结果证实了所提方法在计算复杂度大大降低的前提下达到与最新深度学习的手势识别方法相近性能甚至在复杂数据集上获得最优性能。

参考文献

[1] MITRA S, ACHARYA T. Gesture recognition: A survey[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2007, 37(3): 311-324.

[2] YASEN M, JUSOH S. A systematic review on hand gesture recognition techniques, challenges and applications [J]. PeerJ Computer Science, 2019, 5: e218.

[3] DING I, HSIEH M C. A hand gesture action-based emotion recognition system by 3D image sensor

information derived from Leap Motion sensors for the specific group with restlessness emotion problems[J]. Microsystem Technologies, 2022, 28(1): 403-415.
[4] DING I J, TSAI C Y, YEN C Y. A design on recommendations of sensor development platforms with different sensor modalities for making gesture biometrics-based service applications of the specific group[J]. Microsystem Technologies, 2022, 28(1): 153-166.
[5] 王子懿,沈三民,佘硕铖.基于平面电容传感器阵列的动态手势识别技术[J].测试技术学报,2023,37(1):54-59.
[6] 张瑞轩,张绪树,郭媛,等.基于表面肌电信号的手势识别与分析[J].医用生物力学,2022,37(5):818-825.
[7] 程换新,成凯,程力,等.基于残差融合双流图卷积网络的手势识别方法[J].电子测量技术,2022,45(9):20-24.
[8] OUDAH M, AL-NAJI A, CHAHL J. Hand gesture recognition based on computer vision: A review of techniques[J]. Journal of Imaging, 2020, 6(8): 73.
[9] 孙进,张道周,张洋,等.基于双通道空洞卷积神经网络的手势识别[J].传感器与微系统,2022,41(3):126-128.
[10] JAIN D K, MAHANTI A, SHAMSOLMOALI P, et al. Deep neural learning techniques with long short-term memory for gesture recognition[J]. Neural Computing and Applications, 2020, 32(20): 16073-16089.
[11] LI J, HUAI H, GAO J, et al. Spatial-temporal dynamic hand gesture recognition via hybrid deep learning model[J]. Journal on Multimodal User Interfaces, 2019, 13(4): 363-371.
[12] MOHAMMED A A Q, LV J, ISLAM M S, et al. Multi-model ensemble gesture recognition network for high-accuracy dynamic hand gesture recognition[J]. Journal of Ambient Intelligence and Humanized Computing, 2022: 1-14, DOI: 10.1007/s12652-021-03546-6.
[13] 牛雅睿,武一,孙昆,等.基于轻量级卷积神经网络的手势识别检测[J].电子测量技术,2022,45(4):91-98.
[14] 杨艳芳,刘蓉,刘明,等.基于深度卷积长短时记忆网络的加速度手势识别[J].电子测量技术,2019,42(21):109-113.
[15] 程淑红,杨镇豪,王唱.多通道融合下的手势识别算法研究及船舶虚拟交互平台设计[J].计量学报,2022,43(7):856-862.
[16] 杨阿妮,常丹华.神经网络与马尔可夫模型的手势识别系统[J].电子测量技术,2010,33(4):60-64.
[17] ZHAN F. Hand gesture recognition with convolution neural networks[C]. IEEE Conference on Information Reuse and Integration for Data Science(IRI), 2019: 295-298.
[18] CHENG W, SUN Y, LI G, et al. Jointly network: A network based on CNN and RBM for gesture recognition[J]. Neural Computing and Applications, 2019, 31(1): 309-323.
[19] WU X Y. A hand gesture recognition algorithm based on DC-CNN[J]. Multimedia Tools and Applications, 2020, 79(13): 9193-9205.
[20] SUN Y, WENG Y, LUO B, et al. Gesture recognition algorithm based on multi - scale feature fusion in RGB - D images[J]. IET Image Processing, 2023, 17(4): 1280-1290.
[21] TAN Y S, LIM K M, LEE C P. Hand gesture recognition via enhanced densely connected convolutional neural network[J]. Expert Systems with Applications, 2021, 175: 114797.
[22] TAN Y S, LIM K M, TEE C, et al. Convolutional neural network with spatial pyramid pooling for hand gesture recognition[J]. Neural Computing and Applications, 2021, 33(10): 5339-5351.
[23] 张重生,陈杰,李岐龙,等.深度对比学习综述[J].自动化学报,2023,49(1):15-39.
[24] 李希,刘喜平,李旺才,等.对比学习研究综述[J].小型微型计算机系统,2023:1-14.
[25] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 9729-9738.
[26] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]. International Conference on Machine Learning, 2020: 1597-1607.
[27] CHEN X, FAN H, GIRSHICK R, et al. Improved baselines with momentum contrastive learning[J]. ArXiv Preprint, 2020, ArXiv: 2003.04297.
[28] AKASH M. Image data set for alphabets in the American Sign Language[DB/OL]. DOI: 10.34740/kaggle/dsv/29550.
[29] KUMAR P P, VADAKKEPAT P, LOH A P. Hand posture and face recognition using a Fuzzy-Rough Approach[J]. International Journal of Humanoid Robotics, 2010, 7(3): 331-356.
[30] HE K, ZHANG X, REN SS, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

作者简介

卞雨玮,本科,主要研究方向为机器视觉。
E-mail:201983460001@nuist.edu.cn

周媛(通信作者),博士,副教授,主要研究方向为机器学习、机器视觉和因果生成等。
E-mail:zhouyuan@nuist.edu.cn