计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(11):111−119 [doi: 10.15888/j.cnki.csa.008770]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于残差双注意力与跨级特征融合模块的静态手势识别①

吴佳璐, 田秋红, 岳金鸿

(浙江理工大学 信息学院, 杭州 310018)
通信作者: 田秋红, E-mail: tianqiuhong@zstu.edu.cn

摘　要: 为解决卷积神经网络提取特征遗漏、手势多特征提取不充分问题, 本文提出基于残差双注意力与跨级特征融合模块的静态手势识别方法. 设计了一种残差双注意力模块, 该模块对 ResNet50 网络提取的低层特征进行增强, 能够有效学习关键信息并更新权重, 提高对高层特征的注意力, 然后由跨级特征融合模块对不同阶段的高低层特征进行融合, 丰富高级特征图中不同层级之间的语义和位置信息, 最后使用全连接层的 Softmax 分类器对手势图像进行分类识别. 本文在 ASL 美国手语数据集上进行实验, 平均准确率为 99.68%, 相比基础 ResNet50 网络准确率提升 2.52%. 结果验证本文方法能充分提取与复用手势特征, 有效提高手势图像的识别精度.

关键词: 手势图像识别; ResNet; 残差双注意力模块; 跨级特征融合; 深度学习

引用格式: 吴佳璐,田秋红,岳金鸿.基于残差双注意力与跨级特征融合模块的静态手势识别.计算机系统应用,2022,31(11):111–119. http://www.c-s-a.org.cn/1003-3254/8770.html

Static Gesture Recognition Based on Residual Double Attention Module and Cross-level Feature Fusion

WU Jia-Lu, TIAN Qiu-Hong, YUE Jin-Hong

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: To solve the problems of missing feature extraction by convolutional neural network and insufficient multi-feature extraction of a gesture, this study proposes a static gesture recognition method based on a residual double attention module and a cross-level feature fusion module. The designed residual double attention module can enhance the low-level features extracted by a ResNet50 network, effectively learn the key information, update the weight, and improve the attention to high-level features. Then, the cross-level feature fusion module fuses the high-level and low-level features in different stages to enrich the semantic and location information between different levels in the high-level feature map. Finally, the Softmax classifier of the fully connected layer is used to classify and recognize the gesture image. The experiment is carried out on the American sign language (ASL) dataset. The average recognition accuracy is 99.68%, which is 2.52% higher than that of the basic ResNet50 network. The results show that the proposed method can fully extract and reuse gesture features and effectively improve the recognition accuracy of gesture images.

Key words: gesture image recognition; ResNet; residual double attention module; cross-level feature fusion module; deep learning

① 基金项目: 国家自然科学基金 (51405448); 浙江理工大学博士科研启动项目 (11122932611817); 国家级大学生创新创业训练计划 (11120032382104); 浙江省大学生科技成果推广项目 (14530031661961); 浙江理工大学信息学院教育教学改革研究项目 (11120033312202)
收稿时间: 2022-02-12; 修改时间: 2022-03-14; 采用时间: 2022-03-28; csa 在线出版时间: 2022-07-07

© 中国科学院软件研究所 http://www.c-s-a.org.cn

我国听力残疾人群居各类残疾人群之首, 却由于手语低普及度与文字交流不便性, 与健听人士之间的沟通障碍仍未得到有效解决. 手语是听障人士在日常交际中最为熟悉的自然语言, 通过手的手形、位置、运动轨迹等表达不同语义. 手势识别作为一个热门研究方向, 可实现手语的可读化, 有效打破聋健交互障碍. 此外, 手势识别也可满足人机自然交互需求, 推动机器控制、体感游戏、虚拟现实等应用领域发展. 但由于手势动作多样性、语义复杂性、时空差异性, 手势识别仍面临着众多挑战.

根据静态手势识别的研究方法, 看可分为基于数据手套的手势识别与基于计算机视觉的手势识别. 基于数据手套的手势识别是利用多种传感器获取手部的不同角度信息、运动轨迹信息和时间信息, 将不同信息融合进行分类. 该方法可获得准确的手部信息, 识别率较高, 但其需依赖大量穿戴式传感器, 人机交互的易用性与自然性较低, 使用成本较高, 日常生活不易普及. 相比之下, 基于计算机视觉的手势识别展现出更明显的优势, 利用相关设备采集手势原始图像, 经过图像处理、特征提取、分类识别等步骤获取手势的语义结果, 其中特征提取至关重要, 直接影响着识别准确率.

传统手势特征算法主要基于人工设计特征, 如利用方向梯度直方图 (histogram of oriented gradient, HOG)[1]、尺度不变特征变换 (scale-invariant feature transform, SIFT)[2] 等特征算子提取特征, 在算法的不断优化下识别准确率普遍可达 90%. 吴晓雨等[3] 通过 Kinect 的深度信息完成人手定位, 而后在定位区域内提取基于梯度方向直方图 (HOG) 的形状特征并利用级联 AdaBoost 训练的手势模型, 实现对静态手势的准确识别. 唐文权等[4] 提出了一种基于 YCbCr 颜色空间的改进三帧差分法的动态手势识别方法, 识别率达到 94.6%, 但该方法仅能提取手势图像的显著特征, 容易造成重要特征的遗漏. 由于手势动作灵活多变、背景环境复杂, 特征信息难以完整获取, 人工设计特征就更为复杂、费时, 往往无法发现具有针对性的最高层特性, 从而造成模型学习能力不足, 也会由于提取的特征误差而造成识别率下降.

近年来, 随着深度学习的迅速发展, 卷积神经网络以自动提取特征、权值共享、输入图像与网络结构结合良好等优势[5], 在一系列图像识别问题中表现出显著效果, 成为各类图像识别的主流趋势. Pigou 等[6] 提出基于卷积神经网络的手语识别方法, 运用 CNN 自动匹配特征构建过程, 高精度地识别 20 个意大利手势, 交叉验证准确率为 91.7%. Garcia 等[7] 提出了一种重点关注手型变化的 CNN 网络用于手语识别, 将手型特征送入一个端到端的弱监督分类框架完成识别, 并能对小规模孤立词手语数据集进行实时识别. Jain 等[8] 提出利用支持向量机 (SVM) 与卷积神经网络 (CNN) 对美国手语进行识别, 通过改变滤波器尺寸提高 CNN 模型的精度. 但简单的 CNN 网络提取特征能力有限, 准确率难以进一步提升. Bheda 等[9] 提出利用深度卷积网络对美国手语中的字母和数字图像进行分类的方法, 丰富特征的提取. Bantupalli 等[10]提出一种基于 Inception 模型的 RNN 网络用于手语识别. Yang 等[11] 使用改进的 MobileNet 作为特征提取器, 与 SSD 框架配合使用, 并采用自顶向下的特征融合体系结构, 较好地实现手部检测. 吴晓凤等[12] 提出基于 Faster R-CNN 的手势识别算法, 修改 Faster R-CNN 框架的关键参数, 达到同时检测和识别手势的目的, 并进一步提高识别准确率, 但在复杂背景中的手势识别效果较差. 陈影柔等[13] 提出了一种基于多特征加权融合的静态手势识别方法, 将分割后的手势图像的局部特征与手势灰度图像的深层次特征进行加权融合, 该方法具有较高的识别精度, 但计算量很大. 深层卷积神经网络能够有效提取出更深层次的手势语义信息, 但往往需要较大计算量且伴随着过拟合风险, 难以保证手势识别的高效性与准确性. 并且随着网络不断加深, 特征的感受野明显不足, 会遗漏低层网络提取的手势特征, 不同层级的语义信息无法进行充分复用. 而静态手势识别存在着特征复杂多样、手指间距离角度多变、复杂背景环境干扰等难题, 需要丰富特征多样性, 增强对判别特征的关注, 排除背景特征的干扰.

针对上述问题, 本文提出了基于残差双注意力与跨级特征融合模块的静态手势识别方法, 其主要贡献如下: (1) 设计一个残差双注意力模块, 对特征图不同通道与空间赋予不同权值, 为重要特征分配高注意力, 使网络关注更特定的高层特征, 减轻复杂背景的干扰. (2) 引入轻量的通道注意力与空间注意力模块, 降低网络过深、参数过多带来的过拟合风险, 减少网络的冗余, 加快收敛速度. (3) 增加跨级融合, 将低层特征与不同阶段的高层特征进行融合, 提高特征复用率, 有效丰富不同层次判别特征.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

1 网络结构

1.1 基于残差双注意力与跨级特征融合的手势识别

为解决卷积神经网络提取特征遗漏、手势多特征提取不充分问题, 本文提出基于残差双注意力与跨级特征融合的手势识别方法, 该方法首先将预处理后的原始手势图像输入 ResNet50 基础网络提取低层特征提供充足的感受野, 给出低层特征图, 然后通过改进的残差双注意力模块, 从通道与空间两个层面为不同分辨率特征信息分配权重, 使网络更专注于有目标的判别特征, 高效完成高层特征的提取, 再将不同阶段的高低层特征附加恒等映射关系实现跨级特征融合, 提高特征复用率, 增益不同层次的判别特征. 完成多层特征融合后, 先对特征张量进行全局平均池化, 防止过拟合, 增强特征与类别上映射关系, 最后将特征张量输入到由全连接层和 Softmax 层组成的分类器中完成手势图像的分类识别. 具体流程如图 1 所示.

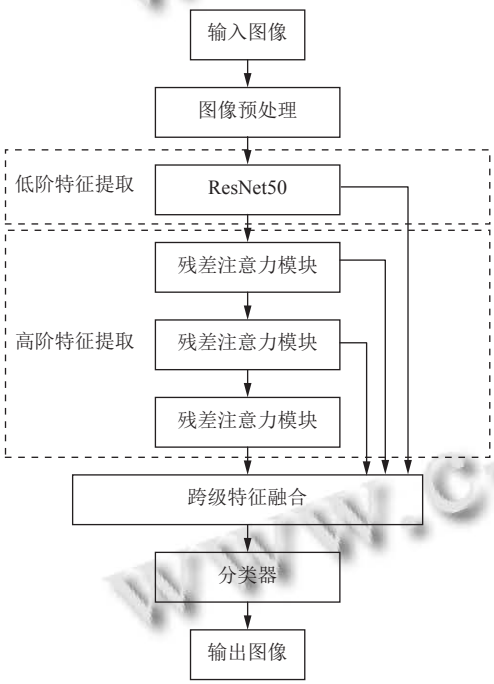

图 1 手势识别流程图

1.2 基于 ResNet50 的低层特征提取

手势图像特征具有高维、多样、强背景干扰等性质, 浅层网络难以充分留存大量有效特征信息. 由此, 本文采用 ResNet50 残差网络作为低层特征提取网络, 提供充足感受野, 提取丰富的关键特征, 输出低层特征图. ResNet50 网络结构如图 2 所示, 主要将两种残差模块 ID block 与 Conv block 加入进传统卷积网络中, 对原始手势图像进行低阶特征提取. 残差结构由 Conv 卷积层、Batch 归一化层或 ReLU 层组成, 其中 Batch 归一化层可较好解决网络过深带来的梯度消失与退化问题, 促使网络的性能可随着层数的叠加而提升, ReLU 层可提升网络非线性表达能力, 对输入图像的低层特征进行有效提取, 同时引入跳跃连接将若干个卷积层前的输入直接与卷积层学习到的特征进行叠加, 简化学习目标同时保证特征信息完整度.

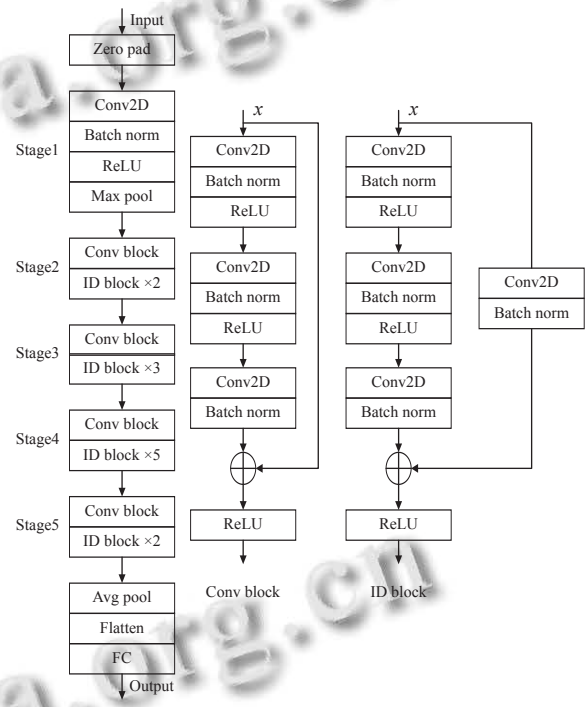

图 2 ResNet50 网络结构图

1.3 基于改进的残差双注意力的高层特征提取

原始输入手势图像通过 ResNet50 基础网络后获得的低层特征图中保留了手势的大部分轮廓与纹理等显著特征, 但其中仍存在大量背景、光照带来的干扰信息, 影响分类的准确率, 需要构建更深层次的卷积神经网络, 增强对判别特征的关注, 排除背景特征的干扰. 故本文设计基于残差双注意力的高层特征提取模块, 在保留特征图原始尺寸的基础上从通道与空间两个层面为不同分辨率特征信息分配权重, 使网络更专注于有目标的判别特征, 忽视背景特征的影响, 为避免过拟合与网络退化风险, 采用残差结构进行搭建. 具体结构如图 3 所示.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

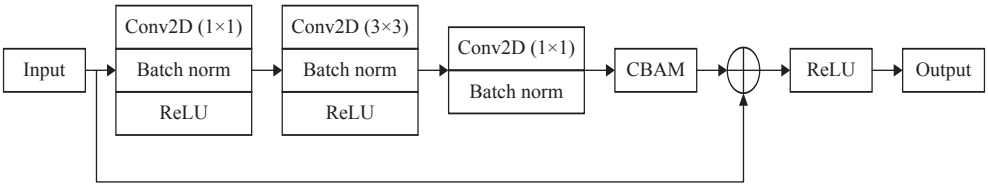

图 3 残差双注意力模块结构图

首先利用 1×1 的卷积核对输入特征图进行降维[14], 简化特征, 然后经过 3×3 的卷积遍历特征图上所有信息, 充分提取特征, 再使用 1×1 的卷积层进行还原并再次挑选有效特征信息, 避免输入信息的损失. 但经过多层卷积操作后会混合空间和通道上的特征, 故引入轻量化的 CBAM[15] 双注意力模块剥离这种混合, 对特征进行有效学习后可灵活更新各通道层面的特征权重, 并学习到空间层面上手势目标像素的关联性, 得到较高层特征图. 为避免纬度上的特征遗漏, 将高层特征与输入特征进行相加融合, 并通过 ReLU 激活函数输出, 增强网络学习能力, 提升计算速度与网络收敛能力. 计算公式如式 (1) 所示:

$$H(x) = F(x, w) + x \tag{1}$$

其中, x 代表输入手势图像的特征张量, $H(x)$ 代表残差双注意力模块的输出, $F(x, w)$ 代表残差双注意力模块中的卷积操作, w 代表权重矩阵.

本文所采用的双注意力模块结构如图 4 所示, 它是一种轻量的前馈卷积神经网络注意力模块, 针对一个中间特征图, 依次通过通道与空间两个独立的维度推断出注意力映射, 然后与原特征图进行乘法加权对特征进行自适应细化[16], 实现增强有效特征, 抑制无效特征的目的, 提升模型的表现力, 特征图经过双注意力模块后的输出形式如下:

$$F1 = Mc(F) \otimes F \tag{2}$$

$$F2 = Ms(F) \otimes F1 \tag{3}$$

其中, F 表示经过 1 个残差模块后的输出, $Mc(\cdot)$ 表示通道注意力映射输出, $Ms(\cdot)$ 表示空间注意力映射输出, $\otimes$表示逐元素相乘.

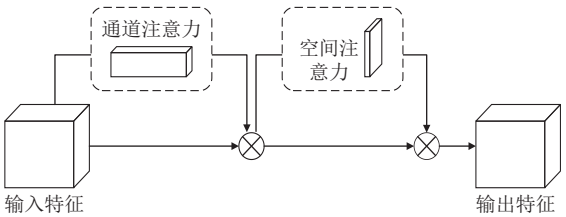

图 4 CBAM 双注意力模块结构图

(1) 通道注意力

持续的卷积操作会导致特征图分辨率大幅度变小, 并且不同通道的信息简单进行融合会造成分割精度下降. 输入手势图像经过 ResNet50 网络提取特征后, 特征图中包含 2048 个通道, 每个通道均提供图像不同特征, 例如形状、线条、空间关系等, 不同特征对图像分类所做贡献也大不相同, 故本文使用通道注意力使网络关注不同通道之间的特征关系, 自主学习一组权重系数, 然后动态加权到每个特征通道上. 通过赋予每个特征通道不同的权重, 使网络自动学习到不同特征通道的重要程度, 更加关注对网络有意义的通道, 从而突出判别特征, 抑制次要特征.

传统通道注意力网络 SENet[17] 首先通过一个空间的全局平均池化进行压缩, 经过两个全连接层和激活函数, 得到通道各自的权重矩阵. 由于全连接层参数较多, 加重网络负担, 增加网络过拟合风险[18], 因此本文对通道注意力模块进行轻量化, 结构如图 5 所示.

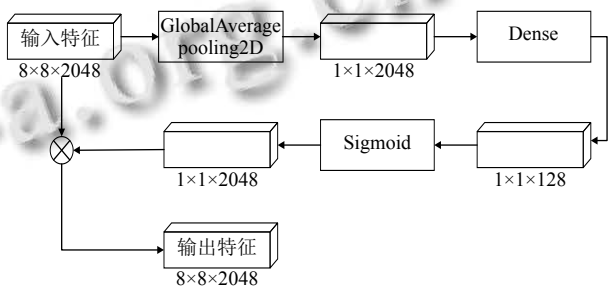

图 5 通道注意力模块结构图

F 为输入特征, 维度为 8×8×2048, 首先将输入特征图经过全局平均池化, 对空间信息进行压缩, 维度压缩至 1×1×2048, 输入包含一个隐藏层的多层感知机 (MLP), 采用一维自适应卷积层对特征通道信息进行处理, 最后使用 Sigmoid 激活函数学习通道间的相关性, 赋予每个特征通道不同权重, 使得网络对各个通道的特征更有辨别能力. 其过程表示为:

$$Mc(F) = \sigma(MLP(Avgpool(F))) \tag{4}$$

© 中国科学院软件研究所 http://www.c-s-a.org.cn

其中, F 表示输入, $Mc(\cdot)$ 表示通道注意力输出, $Avgpool(\cdot)$ 表示全局平均池化操作, $MLP(\cdot)$ 表示全连接层, σ 表示 Sigmoid 激活函数.

(2) 空间注意力

空间注意力模块是对通道注意力的补充, 弥补通道注意力无法充分捕捉图像特定目标位置信息的缺陷. 不是手势图像中所有的区域对分类的贡献都是同等重要的, 低层特征图仅仅映射原始输入图像, 将注意力集中在手部区域, 减少背景信息提供的干扰信息, 能够有效提高手势图像分类性. 因此, 本文采用一种轻量的空间注意力模块, 其结构如图 6 所示.

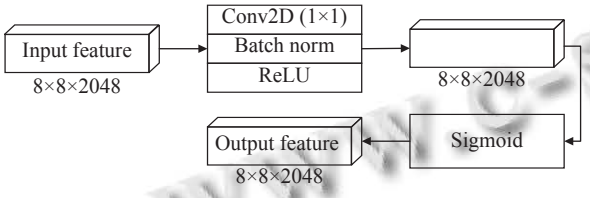

图 6 空间注意力模块结构图

F 为输入特征, 维度为 8×8×2048, 经过自适应的 1×1 卷积操作维度保持不变, 能够挖掘到各个位置信息之间的内在联系, 然后通过 Sigmoid 激活函数生成不同空间注意权重, 从而使得网络更加关注存在手势目标的区域. 其过程表示为:

$$Ms(F) = \sigma(Conv_{1\times1}(F)) \tag{5}$$

其中, F 表示输入, $Ms(\cdot)$ 表示空间注意力输出, $Conv_{1\times1}(\cdot)$ 表示 1×1 的卷积运算, σ 表示 Sigmoid 激活函数.

1.4 跨级特征融合

高层特征增强对判别特征的关注, 排除背景特征的干扰, 但经过深层网络的提取之后容易丢失较多的手势细节信息, 而低层特征保留了手势的大部分轮廓与纹理等细节特征, 但由于提取时的层次较浅, 特征表达的能力不强. 由此, 本文在主干网络中的每个模块后都加入侧向输出用于提供不同层次手势特征信息, 其中 ResNet50 侧向输出弱语义性的低层特征, 各层残差注意力模块侧向输出较高语义性的高层特征. 将不同尺度的高低层特征附加密集连接融合, 使获取的高级特征图聚合更广泛的语义与位置信息, 提高特征的复用率, 减少不同层次信息的丢失.

本文设计的跨级特征融合模块如图 7 所示, 低层网络提取的特征直接输入通过残差双注意力模块的高层特征提取网络, 之后增加一个反向网络, 对高层特征进行 1×1 卷积后与对应的上层特征进行融合, 由此形成 4 层尺度不同的特征 $\{C_1, C_2, C_3, C_4\}$, 最后通过 *Concat* 融合方式融合这 4 层特征, 加强高低层特征的联系, 使最终输出特征图能够融合各层语义、位置信息. 其中 1×1 卷积运算用来维持原有维度, 并进行通道匹配. 跨级特征融合模块过程如下所示:

$$C_1 = Conv_{1\times1}(G_2) \tag{6}$$

$$C_2 = Add(Conv_{1\times1}(G_1), C_1) \tag{7}$$

$$C_3 = Add(Conv_{1\times1}(G_0), C_2) \tag{8}$$

$$C_4 = Add(Conv_{1\times1}(F_0), C_3) \tag{9}$$

$$F_{FG} = Concat(C_1, C_2, C_3, C_4) \tag{10}$$

其中, F_0, G_0, G_1, G_2 表示各层特征提取模块的输出, F_{FG} 表示跨级融合后的输出, $Conv_{1\times1}(\cdot)$ 表示 1×1 的卷积运算, $Add(\cdot)$ 表示特征求和操作, $Concat(\cdot)$ 表示特征融合操作.

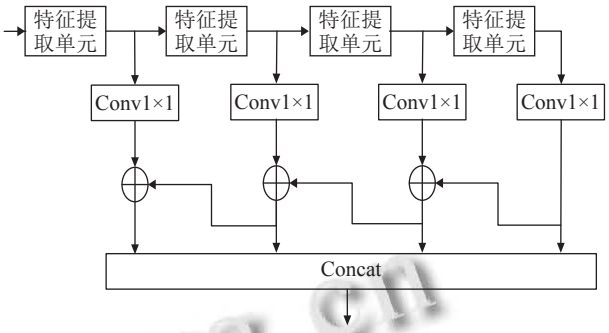

图 7 跨级特征融合模块结构图

2 实验结果与分析

2.1 实验设置

(1) 实验环境

本文实验均是在 macOS 11.3.1 系统下、采用基于 Python 的 TensorFlow 深度学习框架进行的. 处理器为 2.3 GHz 双核 Intel Core i5, 显卡为 Nvidia Tesla V100 (32 GB), TensorFlow 版本为 2.5.0, 编译环境为 Python 3.7.3.

(2) 实验超参数设置

在训练过程中, 通过大量实验选择最优超参数, 以提高网络的拟合性能. 最终选用 Adam 算法作为模型参数优化器, 设置初始学习率为 0.000 01, 批处理大小为 16, 共迭代 300 次, 采用迁移学习方法训练网络模型.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

(3) 实验数据

在实验中, 选用 ASL 美国手语数据集与 BSL 孟加拉手语数据集进行训练. ASL 数据集为美国手语字母图像的集合, 包含 29 类手势, 其中 26 类为字母 A–Z, 其余 3 类为 space、del、nothing, 其手势图像背景复杂, 存在与肤色相近的颜色, 且拍摄亮度、距离各有不同. 通过 ASL 数据集重点测试本文模型在手势图像背景复杂情况下的识别效果. 随机选择其中 6 000 张手语图像, 按照 4:1 的比例划分训练集和测试集, 每幅图像对像素大小为 200×200, 部分手语图像如图 8 所示. BSL 数据集为孟加拉手语数字图像集合, 包含 38 类手势, 为数字 0–37, 其手势特征复杂多样, 手部轮廓、手指间距、手指弯曲程度等都相似度较高. 通过 BSL 数据集重点测试本文模型在手势特征复杂多样情况下的识别效果. 随机选择其中 10 000 张手语图像, 按照 4:1 的比例划分训练集和测试集, 每幅图像对像素大小为 224×224, 部分手语图像如图 9 所示.

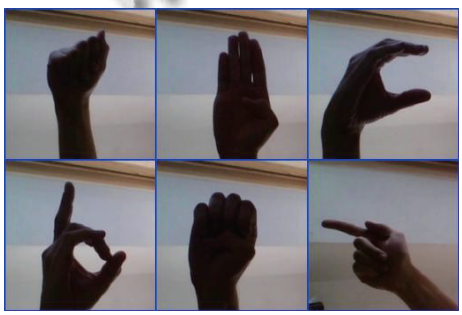

图 8 ASL 手语数据集部分手语图像

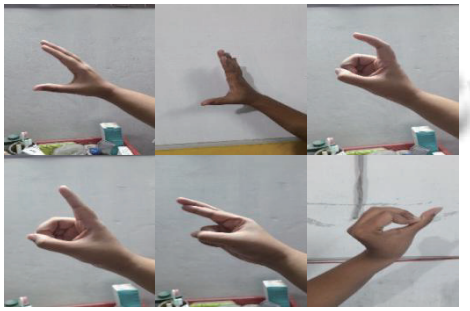

图 9 BSL 手语数据集部分手语图像

为减少计算量并提升训练速度, 在训练前先对手势图像进行预处理, 将原始手势图像尺寸归一化为 256×256 的 RGB 三通道图像, 再对 RGB 图像进行标准化, 将 [0, 255] 之间的整数缩小为 [0, 1] 之间的浮点数.

(4) 评估指标

为使实验结果更为可靠, 本文在经过随机乱序后的验证集上执行 5 次, 取得平均识别准确率作为训练结果, 准确率是指对于给定数据集, 在不考虑样本实际类别的情况下, 正确分类识别的样本数占所有样本数的比例: $Acc=(N_{correct})/(N_{total})$, $N_{correct}$ 表示验证集中分类识别结果正确的样本数, N_{total} 表示验证集样本总数.

2.2 不同基础网络性能实验对比

为验证所采用的 ResNet50 基础网络模型是最适合且有效的, 本文将 VGG16、VGG19、Xception、ResNet50、ResNet152 这 4 种网络分别作为基础网络进行对比分析. 采用 ASL 数据集作为训练集和测试集, 为每个模型训练设置相同实验参数, 并确保其在相同条件下完成训练, 直至收敛.

图 10 是不同网络在 ASL 数据集上准确率随迭代次数的变化曲线图, 由图可知, ResNet50 与 Xception 网络的识别效果明显优于其他网络, 准确率上升速度最快, 迭代 10 次后准确率就能达到 97%, 之后准确率便趋于稳定, 相较于其他网络模型, 获得更高的识别准确率, 而 ResNet50 最终的识别准确率较高于 Xception.

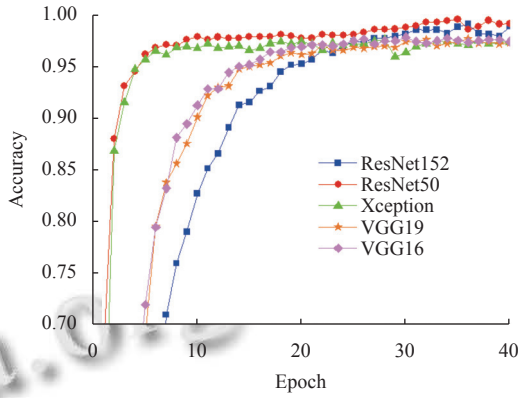

图 10 不同基础网络准确率随迭代次数变化曲线图

图 11 是不同网络在 ASL 数据集上损失值随迭代次数的变化曲线图, 由图 11 可知, VGG16、VGG19、ResNet152 网络均迭代 20 次后才进入稳定收敛, 而 ResNet50 与 Xception 网络只需迭代 10 次, 可见 ResNet50 与 Xception 网络的收敛效果更好, 而 ResNet50 的最终损失值最小, Xception 的最终损失值较大, 综上 ResNet50 网络更为优异.

模型的参数量、每轮平均训练时间与平均识别准确率比较如表 1 所示, VGG16 与 VGG19 的每个 epoch 的平均训练时间较短. 但其识别准确未能突破 98%, ResNet 系列网络模型运行时间加长, 是 VGG 系列的 1.5 倍及以上, 但准确率均在 98% 以上, 其中 ResNet152

© 中国科学院软件研究所 http://www.c-s-a.org.cn

网络层数更深, 每轮训练时间几乎是 ResNet50 网络的 1.3 倍, 但识别准确率却下降了 1.57%. 综上所述, ResNet50 网络能节省特征提取与学习特征的时间, 在较短的运行时间内实现较高的识别准确率, 具有优异的识别效果.

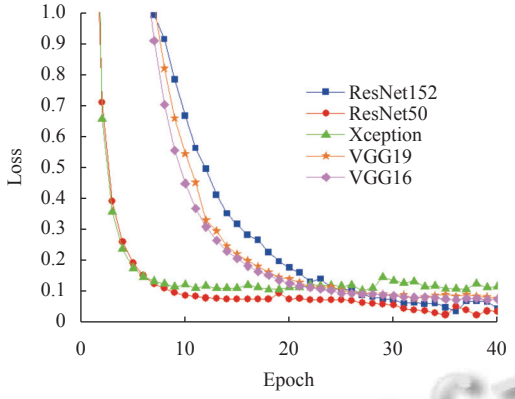

图 11 不同基础网络损失值随迭代次数变化曲线图

表 1 不同基础网络性能实验结果对照

模型	参数量 (M)	每轮平均训练时间 (s)	准确率 (%)
VGG16	19.4	24	97.67
VGG19	24.7	26	97.50
Xception	90.8	39	97.58
ResNet50	93.5	36	99.68
ResNet152	128.3	48	98.21

2.3 不同残差注意力块个数性能实验对比

为选择最为合适的网络层数, 使网络识别精度达到最优, 本文保持其他参数设置不变, 依次选择不同残差注意力块个数进行实验, 初始学习率设置为 0.000 01, 在训练集上一次迭代训练 40 个 epoches 后, 每个 epoch 包含 375 张手语图像. 不同个数残差注意力块所对应网络的参数量、每轮平均训练时间与平均识别准确率如表 2 所示.

表 2 不同残差注意力块个数性能实验结果对照

残差注意力块个数	参数量 (M)	每轮平均训练时间 (s)	准确率 (%)
0	24.8	13	97.16
1	50.5	20	98.21
2	72.0	29	98.28
3	93.5	36	99.68
4	115.1	41	98.00
5	136.6	47	97.90

由表 2 可知, 增加残差注意力块个数时会增加网络的总体参数量, 使网络的训练速度减慢, 但其对网络性能的提升有明显效果. 结果表明, 当残差注意力块个数为 3 时, 网络的性能最优, 准确率达到 99.68%, 说明网络能充分利用学习到的特征通过分类识别的精度. 当继续增加残差注意力块时, 网络的拟合能力达到饱和, 导致识别精度下降. 故本文选择残差注意力块个数为 3.

2.4 消融实验对比

本文方法核心在于残差注意力模块与跨级特征融合, 为验证各模块对整体网络模型的效能, 本节在相同其余实验参数情况下, 基于 ASL 美国手语数据集进行多组消融实验, 以说明本文网络中所有设置均为最优. 实验以 ResNet50 为基础网络, 在基础网络上添加残差模块、通道注意力机制、空间注意力机制、CBAM、跨级特征融合, 初始学习率设置为 0.000 01, 在训练集上一次迭代训练 40 epoches 后, 实验结果如表 3 所示. 其中, 模型 1 表示基础的 ResNet50, 模型 2 加入通道注意力, 模型 3 加入空间注意力, 模型 4 采用 CBAM 模块, 模型 5 在含有 CBAM 模块的网络中引入残差模块, 本文模型增加跨级特征融合模块.

表 3 消融实验结果对照

模型	ResNet50	通道注意力	空间注意力	CBAM	残差模块	跨级特征融合	准确率 (%)
1	√	—	—	—	—	—	97.16
2	√	√	—	—	—	—	97.59
3	√	—	√	—	—	—	97.93
4	√	√	√	√	—	—	98.26
5	√	√	√	√	√	—	98.40
本文	√	√	√	√	√	√	99.68

模型 2、模型 3 相比模型 1 基础的 ResNet50 网络识别准确率分别提升了 0.43% 与 0.77%, 这是注意力模块中的自适应权重更新带来的效益, 能够增强对判别特征的关注, 而空间注意力的效益大于通道注意力, 是因为空间注意力模块将目标更专注于手势部分的重要特征, 减少复杂背景的干扰. 模型 4 同时引入两种注意力机制, 准确率提升了 1.08%, 双注意力机制能使整体网络学习到更多目标区域的特征. 模型 5 在含有 CBAM 模块的网络中引入残差模块, 使准确率提升了 0.14%. 本文模型增加跨级特征融合模块使得准确率提升 1.28%, 可见其在分类识别中的显著效果, 能够充分利用不同阶段提取的高低层判别特征. 糅合残差注意力模块与跨级特征融合, 可达到最优的识别效果, 相比模型 1 基础 ResNet50 网络准确率提升 2.52%, 这有效验证本方法在所有设置实现最优. 实验证明, 通过深层卷积神经网络对原始输入图像进行特征提取与描述过

© 中国科学院软件研究所 http://www.c-s-a.org.cn

程中, 易于受到背景、光照等特征干扰, 丢失某些有用的语义特征, 无法保证获得满意的分类识别结果. 通过基于残差双注意力与跨级特征融合模块的手势识别方法, 很大程度上提高判别特征提取与复用, 促进识别准确率的提升.

2.5 现有手势识别方法对比

为验证提出方法的先进性与普适性, 本文将在 ASL 美国手语数据集上实验的 5 种手势识别方法与在 BSL 孟加拉手语数据集上实验的 5 种手势识别方法分别与本文提出方法进行对比, 具体结果如表 4、表 5 所示, 可知相较于现有最新的手势识别方法, 本文提出的方法在识别准确率上有明显的提升, 这进一步验证本文手势识别方法的先进性, 不仅深层特征相较于其他方法更为丰富, 特征复用能力也优于其他方法. 并且本文提出的方法在 ASL 数据集与 BSL 数据集上准确率均能达到 99% 以上, 可见本文手势识别方法的普适性, 面对手势特征复杂多样、手指间距离角度多变、复杂背景环境干扰等难题, 均能展现出优异的分类识别能力.

表 4 ASL 数据集上现有手势识别方法准确率对比

方法	准确率 (%)	年份
文献[19]	93.81	2018
文献[20]	95.00	2019
文献[21]	99.44	2020
文献[22]	98.07	2021
文献[23]	96.30	2021
本文	99.68	2022

表 5 BSL 数据集上现有手势识别方法准确率对比

方法	准确率 (%)	年份
文献[24]	98.66	2018
文献[25]	96.12	2020
文献[26]	99.22	2020
文献[27]	98.75	2020
文献[28]	98.75	2020
本文	99.62	2022

3 结论与展望

本文针对现有卷积神经网络提取特征遗漏、手势多特征提取不充分问题, 提出了基于残差双注意力与跨级特征融合模块的静态手势识别方法. 该方法使用 ResNet50 网络提取低层手势特征, 并设计基于残差双注意力的高层特征提取模块, 从通道与空间两个层面为不同分辨率特征信息分配权重, 增益判别特征, 抑制背景特征, 然后将不同层次的高低层特征进行跨级特征融合, 提高特征的复用率, 减少特征遗漏. 实验结果表明, 与现有方法相比, 本文提出的静态手势识别方法在 ASL 美国手语数据集上的识别效果更具先进性与有效性. 但本文手势识别研究局限于静态图像, 而现实生活中的手势交互必然是动态连续的, 因此未来工作将研究动态手势识别, 致力于设计出高效、准确的用于动态手势识别的网络模型.

参考文献

1 Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego: IEEE, 2005. 886–893.

2 Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

3 吴晓雨, 杨成, 冯琦. 基于 Kinect 的手势识别算法研究及应用. 计算机应用与软件, 2015, 32(7): 173–176, 276. [doi: 10.3969/j.issn.1000-386x.2015.07.042]

4 唐文权, 徐武, 文聪, 等. 复杂背景下基于肤色检测的动态手势分割与识别. 科学技术与工程, 2019, 19(33): 330–335. [doi: 10.3969/j.issn.1671-1815.2019.33.049]

5 李航, 厉丹, 朱晨, 等. 基于卷积神经网络的图像识别系统. 电脑知识与技术, 2020, 16(10): 196–197, 200.

6 Pigou L, Dieleman S, Kindermans PJ, *et al*. Sign language recognition using convolutional neural networks. Proceedings of ECCV 2014 Workshops. Cham: Springer, 2015. 915–922.

7 Garcia B, Viesca SA. Real-time American sign language recognition with convolutional neural networks. Convolutional Neural Networks for Visual Recognition, 2016, 2: 225–232.

8 Jain V, Jain A, Chauhan A, *et al*. American sign language recognition using support vector machine and convolutional neural network. International Journal of Information Technology, 2021, 13(3): 1193–1200. [doi: 10.1007/s41870-021-00617-x]

9 Bheda V, Radpour D. Using deep convolutional networks for gesture recognition in American sign language. arXiv: 1710.06836, 2017.

10 Bantupalli K, Xie Y. American sign language recognition using deep learning and computer vision. Proceedings of the 2018 IEEE International Conference on Big Data (Big Data).

© 中国科学院软件研究所 http://www.c-s-a.org.cn

Seattle: IEEE, 2018. 4896–4899.

11 Yang L, Qi Z, Liu ZH, *et al*. An embedded implementation of CNN-based hand detection and orientation estimation algorithm. Machine Vision and Applications, 2019, 30(6): 1071–1082. [doi: 10.1007/s00138-019-01038-4]

12 吴晓凤, 张江鑫, 徐欣晨. 基于 Faster R-CNN 的手势识别算法. 计算机辅助设计与图形学学报, 2018, 30(3): 468–476.

13 陈影柔, 田秋红, 杨慧敏, 等. 基于多特征加权融合的静态手势识别. 计算机系统应用, 2021, 30(2): 20–27. [doi: 10.15888/j.cnki.csa.007748]

14 赵文清, 孔子旭, 周震东, 等. 增强小目标特征的航空遥感目标检测. 中国图象图形学报, 2021, 26(3): 644–653. [doi: 10.11834/jig.190612]

15 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of Computer Vision—ECCV 2018. Cham: Springer, 2018. 3–19.

16 吴若有, 王德兴, 袁红春, 等. 基于多分支全卷积神经网络的低照度图像增强. 激光与光电子学进展, 2020, 57(14): 141021.

17 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

18 宋东情, 朱定局, 贺超. 基于多模型特征与精简注意力融合的图像分类. 计算机系统应用, 2021, 30(11): 210–216. [doi: 10.15888/j.cnki.csa.008153]

19 Chong TW, Lee BG. American sign language recognition using leap motion controller with machine learning approach. Sensors, 2018, 18(10): 3554. [doi: 10.3390/s18103554]

20 Bin LY, Huann GY, Yun LK. Study of convolutional neural network in recognizing static American sign language. Proceedings of 2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). Kuala Lumpur: IEEE, 2019. 41–45.

21 Lee CKM, Ng KKH, Chen CH, *et al*. American sign language recognition and training method with recurrent neural network. Expert Systems with Applications, 2021, 167: 114403. [doi: 10.1016/j.eswa.2020.114403]

22 Rivera-Acosta M, Ruiz-Varela JM, Ortega-Cisneros S, *et al*. Spelling correction real-time American sign language alphabet translation system based on YOLO network and LSTM. Electronics, 2021, 10(9): 1035. [doi: 10.3390/electronics10091035]

23 Sharma S, Kumar K. ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks. Multimedia Tools and Applications, 2021, 80(17): 26319–26331. [doi: 10.1007/s11042-021-10768-5]

24 Purkaystha B, Datta T, Islam MS. Bengali handwritten character recognition using deep convolutional neural network. Proceedings of 2017 20th International Conference of Computer and Information Technology (ICCIT). Dhaka: IEEE, 2017. 1–5.

25 Ahmed S, Tabsun F, Reyadh AS, *et al*. Bengali handwritten alphabet recognition using deep convolutional neural network. Proceedings of 2019 5th International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). Rajshahi: IEEE, 2019. 1–4.

26 Chatterjee S, Dutta RK, Ganguly D, *et al*. Bengali handwritten character classification using transfer learning on deep convolutional network. Proceedings of 11th International Conference on Intelligent Human Computer Interaction. Cham: Springer, 2020. 138–148.

27 Hasan M, Srizon AY, Hasan AM. Classification of Bengali sign language characters by applying a novel deep convolutional neural network. Proceedings of 2020 IEEE Region 10 Symposium (TENSYMP). Dhaka: IEEE, 2020. 1303–1306.

28 Hossain S, Sarma D, Mittra T, *et al*. Bengali hand sign gestures recognition using convolutional neural network. Proceedings of 2020 2nd International Conference on Inventive Research in Computing Applications (ICIRCA). Coimbatore: IEEE, 2020. 636–641.

(校对责编: 牛欣悦)

© 中国科学院软件研究所 http://www.c-s-a.org.cn