文章编号：2095-2163(2021)06-0051-06 中图分类号：TP391.4 文献标志码：A

基于改进 AlexNet 的双模态握笔手势识别

张 璐，陶 然，彭志飞，丁金洋
（东华大学 计算机科学与技术学院，上海 201600）

摘 要：本文提出了一种基于改进 AlexNet 的双模态握笔手势识别方法。该方法根据握笔手势特征自建了 8 100 张握笔手势数据集，对数据集进行了手势分割获取二值图像、骨架提取获取包含原图的骨架图像等处理，并将处理后的 2 种类型图像构成双模态图像输入至改进的 AlexNet 中。针对 AlexNet 提取握笔手势特征不充分的问题，本文将 AlexNet 第一层的卷积核大小修改为 3×3，并在卷积层之后添加了批量归一化、注意力机制。通过实验证明，该方法对 9 种握笔手势的平均识别率达到 75.6%，分别高于骨架图像、分割图像、AlexNet 网络 11%、16%和 13%，证明了该模型对握笔手势识别的有效性。

关键词：手势分割；骨架提取；双模态输入；AlexNet；握笔手势识别

Bimodal pen-holding gesture recognition based on improved AlexNet

ZHANG Lu, TAO Ran, PENG Zhifei, DING Jinyang
(College of Computer Science and Technology, Donghua University, Shanghai 201600, China)

【Abstract】 In this paper, a new method of pen-holding gesture recognition based on improved AlexNet is proposed. In this method, 8 100 pen-holding gesture data set is constructed according to the characteristics of pen-holding gesture. Then the data set is processed, including gesture segmentation obtains binary image, skeleton extraction obtains the skeleton image containing the original image, and the processed images are formed into bimodal images which are input into the improved AlexNet. In order to solve the problem that AlexNet is not sufficient in extracting the characteristics of pen-holding gesture, this paper modifies the convolution kernel size of AlexNet's first layer to 3×3, meanwhile adds batch normalization and attention mechanism after the convolution layer. The experimental results show that the average recognition rate of nine pen-holding gesture is 75.6%, which is 11%, 16% and 13% higher than that of skeleton image, segmenting image and AlexNet network, respectively. It proves the effectiveness of the proposed model for pen-holding gesture recognition.

【Key words】 gesture segmentation; skeleton extraction; bimodal input; alexNet; pen-holding gesture recognition

0 引 言

近年来，随着计算机视觉应用技术的快速发展，国内外的研究人员对人脸、表情、姿态、手势等人机交互方面进行了大量的研究[1]。较于其它交互方式，手势具有更加直接、灵活、自然的特点，因此手势识别引起了研究者的极大关注[2]。由于当下部分中小学生握笔手势不规范，导致坐姿不标准、眼睛近视以及手指关节增生，对其身心健康造成了不良的影响[3]。

手势分为静态手势以及动态手势识别两种类型，本文仅对静态手势识别进行研究，其关键技术分为手势分割、手势识别两部分。薛俊韬[4]等人利用人体肤色的聚类特性，在 YCbCr 空间构建皮肤颜色分布模型，对手势进行分割，此颜色空间受光照等变化影响较小，肤色的聚类效果好，因此本文手势分割算法基于此颜色空间。谢峥桂等人[5]首先对手势图像进行手势分割预处理，接着对处理后的图像利用 CNN 模型进行特征提取和识别。文献[6]基于卷积神经网络开发了 OpenPose 模型，实现了人体关键点检测以及骨架图的绘制。Mazhar 等人[7]基于 OpenPose 模型构建了手势实时检测人机交互系统。随着深度学习的不断发展，研究者们提出了多模态输入的方法。文献[8]中提出，将骨骼关节信息、深度图像和 RGB 图像同时输入至隐马尔可夫模型的半监督分层动态框架，进行手势分割和识别。

综上所述，本研究受到多模态的启发，提出基于改进 AlexNet 的双模态握笔手势识别方法，将握笔手势分割图像与骨架图像同时输入至增加了批标准化、注意力机制以及修改了卷积核大小的改进

作者简介：张 璐(1997-)，女，硕士研究生，主要研究方向：深度学习；陶 然(1975-)，男，硕士，高级实验师，主要研究方向：大数据、人工智能、数据挖掘；彭志飞(1996-)，男，硕士研究生，主要研究方向：深度学习；丁金洋(1998-)，男，硕士研究生，主要研究方向：深度学习。
通讯作者：陶 然 Email：taoran@dhu.edu.cn

收稿日期：2021-04-12

AlexNet 中,进行手势分割和识别。

1 相关工作

1.1 手势分割

手势分割[9]旨在将图像中手势区域和背景区域分离,从而将手势从图片中提取出来。目前,基于视觉的手势分割方法主要有基于肤色的手势分割方法、基于运动的手势分割方法、基于轮廓的手势分割方法等。由于手势图像是 RGB 形式,光照变化会对肤色分割产生影响,不适合进行肤色分割。而 YCbCr 色彩空间肤色聚类效果好,可将 RGB 图像中的皮肤信息映射到 YCbCr 空间,通过判断某点在 YCbCr 空间的坐标(Cb,Cr)是否在椭圆内,将肤色区域与背景部分区分开。

由于手势分割后的图像包含噪声等,因此利用图像增强技术改善图像的视觉效果,突出图像中计算机感兴趣的部分。图像增强[10]是利用数学形态学对图像进行处理,其中包括图像腐蚀、膨胀、开运算和闭运算等。对图像先腐蚀后膨胀的操作称为开运算,具有分离细小物体的作用。本文使用开运算对握笔手势分割图像进行图像增强,使其进一步优化。

1.2 AlexNet 简介

卷积神经网络(Convolutional Neural Network,CNN)[11]是由 YannLeCun 于 1988 年提出的一种深度前馈神经网络,主要由卷积层、池化层和全连接层组成。AlexNet 是 CNN 模型的历史突破点,之后的网络模型都基于此进行改进。

AlexNet[12]是卷积神经网络最具代表性的模型之一,且在 2012 举行的 ImageNet 大规模视觉识别挑战比赛中获胜。AlexNet 网络由 5 个卷积层、3 个全连接层组成,其网络结构如图 1 所示。

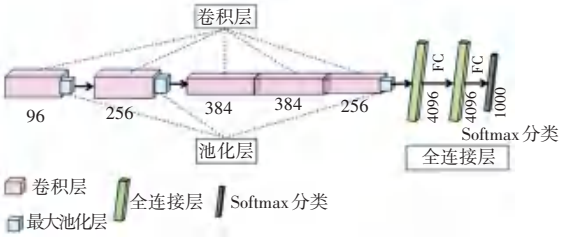

图 1 AlexNet 网络结构图

Fig. 1 AlexNet network structure diagram

AlexNet 相比其它网络具有的优势是:使用了 *ReLu* 激励函数、Dropout、数据增强、最大池化以及局部响应归一化(Local Response Normalization,LRN)技术。

ReLu 函数作为 AlexNet 中的激活函数,有效地防止训练图像识别模型时出现过拟合问题;Dropout 能够使神经元在训练过程中以一定的概率停止,避免了网络模型的过拟合;最大池化解决了平均池化的模糊化问题,丰富了手势图像特征;数据增强通过截取手势图像方式,实现了图像数据量的增加,从而防止过拟合问题的出现,提升网络的泛化能力;LRN 则对当前层的输出结果做平滑处理,增强了网络模型的泛化能力。

2 改进的 AlexNet 双模态握笔手势识别方法

针对单模态卷积神经网络特征提取不充分的问题,本文提出了一种基于改进 AlexNet 的双模态握笔手势识别方法。即将握笔手势分割图像与骨架图像输入至改进的 AlexNet 中进行特征提取、特征融合,最后利用 Softmax 层对 9 类握笔手势进行分类。

2.1 网络结构设计

为了能够更好地解决握笔手势识别问题,本文对 AlexNet 进行了改进。改进的 AlexNet 网络结构如图 2 所示。

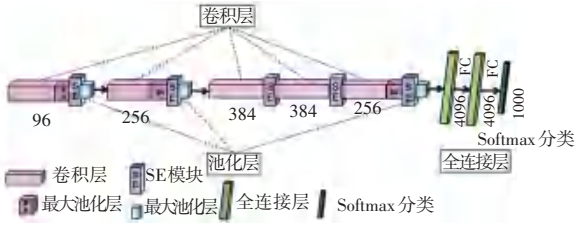

图 2 改进的 AlexNet 网络结构图

Fig. 2 Improved AlexNet network structure diagram

如图 2 所示,本文在卷积层之后添加批量归一化(Batch Normalization,BN)与注意力机制。BN 层用来解决训练过程中识别率出现波动大的问题,注意力机制则用来加强包含握笔手势信息的特征图,添加的注意力机制的结构如图 3 所示。其次,修改了卷积核大小。为了适应1 000种图像的多分类问题,原 AlexNet 网络结构第一个卷积核的大小为 11×11,而本文改进的 AlexNet 是用于 9 种握笔手势识别,因此将 AlexNet 的第一层卷积核大小改为 3×3,不仅能够更好地获取握笔手势图像特征分布,而且可以减少参数训练。本文将 AlexNet 使用的随机梯度下降法替换为自适应时刻估计算法,以自适应调整学习率,减少调参量。本文在有无 LRN 层的模型上进行测试,结果并无区别,因此删除了 LRN 层。

如图 3 所示,SE 模块作用在通道尺度,给不同的通道特征进行加权操作。对于输入的 $C \times H \times W$

的特征图，根据全局平均池化处理得到 C 个标量，然后将输出的结果通过 2 个全连接层以及激励函数得到权重。通过在每个通道的维度上学习、更新不同的权重，最终得到计算注意力的矩阵以加强重要特征。

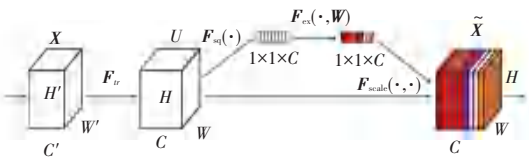

图 3　SE 模块结构图

Fig. 3　Structure of SE Module

2.2　基于改进 AlexNet 的双模态握笔手势识别

由于单模态输入提取特征不丰富，本文提出了双模态输入的方法，即对握笔手势图像进行手势分割以及骨架提取的 2 种处理方式，获取握笔手势分割图像与握笔手势骨架图像。其中骨架提取是在原图上进行的。

在对改进 AlexNet 网络进行模型训练之前，先对握笔手势分割图像以及骨架图像进行数据增强处理，包括：旋转、缩放、平移和尺度变换等；接着对数据集进行尺度归一化，得到 224×224 的图像；最后对处理后的图像，利用改进的 AlexNet 网络进行特征提取、特征融合和手势识别。双模态握笔手势识别框架图如图 4 所示。

图 4　双模态握笔手势识别框架

Fig. 4　Bimodal pen-holding gesture recognition frame diagram

先将双模态握笔手势图像分别输入至改进 AlexNet 网络进行特征提取、特征拼接后，进行全连接操作，利用 Softmax 层进行分类，最后得到握笔手势识别结果。

3　实验结果与分析

3.1　双模态握笔手势数据集的建立

本文在对文献资料[3]的研究基础上，将握笔手势分为 9 种类型，其中包括：标准型、错位型、横搭型、埋头型、扭曲型、扭转型、拳头型、睡觉型和直线型。9 种握笔手势的部分数据集如图 5 所示。

由于握笔手势没有数据集，因此本文严格按照各种类型的标准在不同的角度、背景下自建数据集。每种手势有 900 张，共计 8 100 张。训练集、测试集与验证集以 6∶2∶2 的比例进行划分。接着对握笔手势图像进行分割、骨架提取 2 种处理方式，获取 2 种不同模式的图像，即双模态图像，双模态握笔手势数据集的制作过程如图 6 所示。由于数据集数量的限制，本文对双模态数据集进行数据增强处理。其中包括：旋转、缩放变换、平移变换和尺度变换等，使得握笔手势数据集更加丰富、有效。

首先对握笔手势图像利用颜色空间转换、椭圆肤色模型分割、开运算去噪技术进行握笔手势分割，得到握笔手势分割图像；同时利用 OpenPose 手部模型进行骨架提取，得到握笔手势骨架图像；最后综合得到双模态握笔手势数据集。

3.2　实验与对比

为了评估本算法的优越性，本文对网络参数进行调整后，进行了 3 组对比实验。网络参数首先在改进 AlexNet 网络的基础上对参数进行设置。首先对比了 batch 的大小对网络训练的结果，通过设置 batch 为 16、32、64，得到 3 种识别率的变化，对 3 种识别率进行分析。在 batch 为 64 的情况下，识别率高、收敛速度快且波动小；接着对比了迭代次数为 100 和 150 的情况，结果表明，迭代次数为 150 时，识别率更加稳定；最后对比了 Dropout 系数为 0.5 和 0.8 的情况，选择了 0.5 进行实验，此时识别率波动小，收敛速度较快。在此基础上，本文设置了 3 组对比实验。

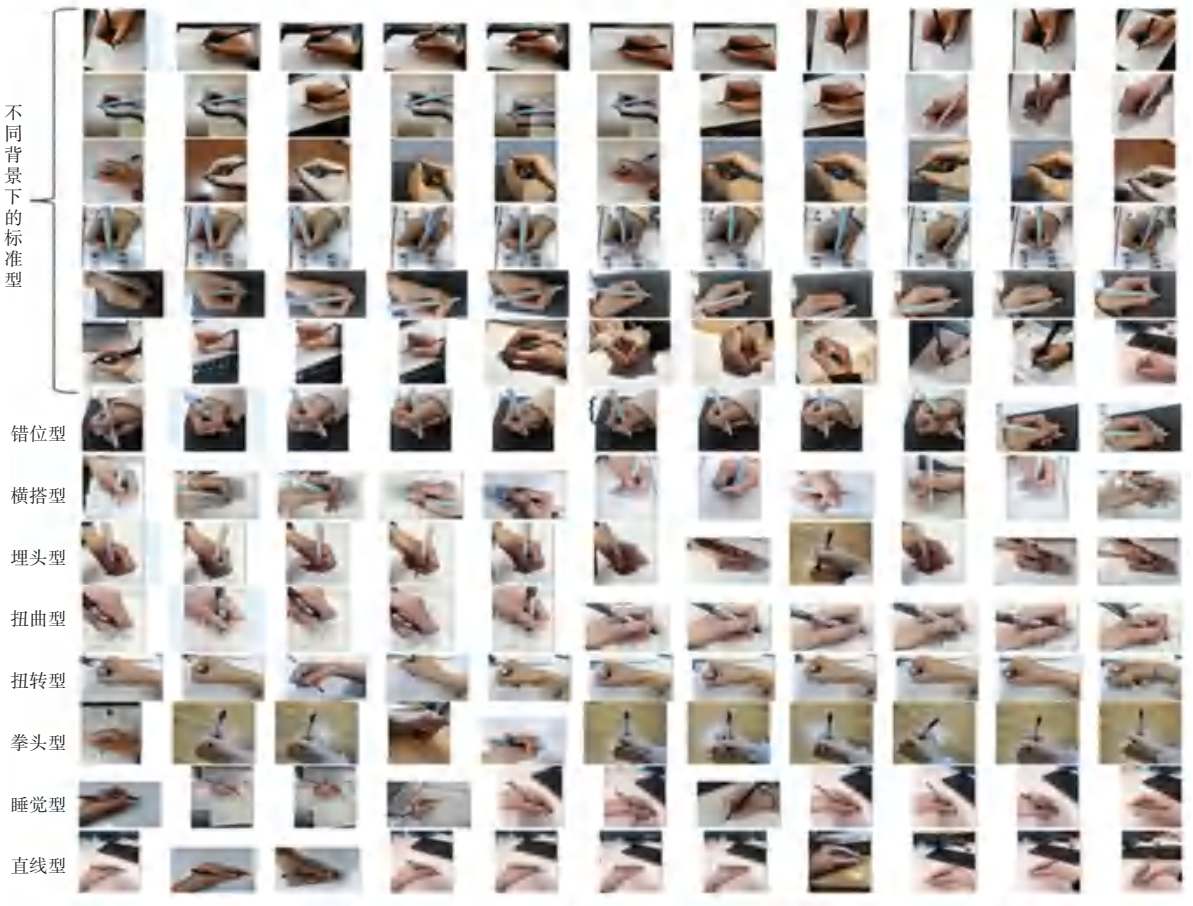

图 5 9 种握笔手势部分数据集展示

Fig. 5 Data set display of nine pen-holding gesture

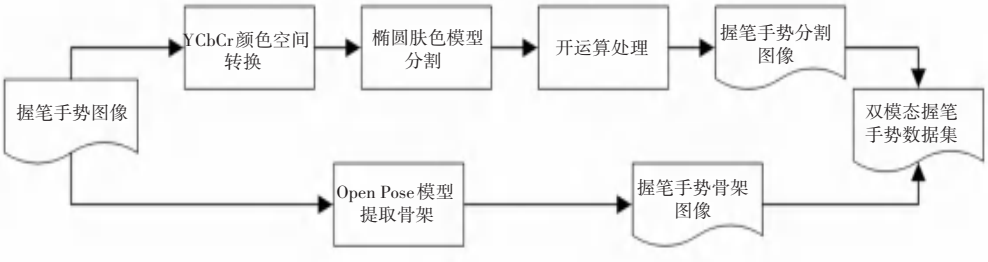

图 6 双模态数据集制作流程

Fig. 6 Production process of bimodal data set

3.2.1 第一组实验

在改进的 AlexNet 中进行。在其它参数不变的情况下,对只添加 BN 层与只添加 SE 模块进行对比,结果如图 7、图 8 所示。

由图 7、图 8 的识别率结果可见,只添加 BN 层的识别率波动小,但是识别率较低;添加了注意力机制的识别率虽然相对于只添加 BN 层的识别率高,但识别率变化起伏波动大。

3.2.2 第二组实验

将握笔手势分割图像、握笔手势骨架图像和双模态图像输入至改进 AlexNet 网络,对识别效果进行比较如图 9 所示。同时,还比较了不同迭代次数下的识别准确率。

通过图 9 可以发现,握笔手势分割图像的识别率比骨架图像、以及双输入图像的识别率低。由于采用握笔手势分割图像进行识别时,手势遮挡使得手指的分割结果不明显;骨架图像尽管因为部分遮挡导致提取不完整,但是因为同一类型的骨架图像提取都有一些缺失,且有原图特征补充,因此骨架图像的识别率比分割图像的识别率高;而双模态图像综合了握笔手势分割图像、骨架图像以及原始图像的特征,使得其识别率高于握笔手势分割图像、骨架图像。

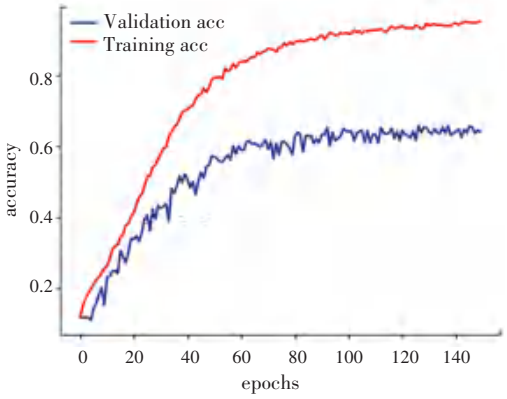

图 7　只添加 BN 层的识别率

Fig. 7　Add only recognition rate of BN layer

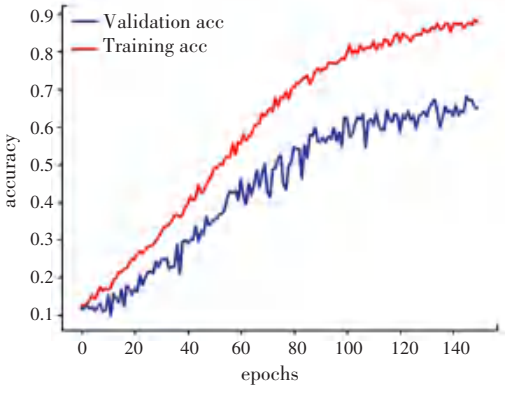

图 8　只添加 SE 模块的识别率

Fig. 8　Add only recognition rate of SE

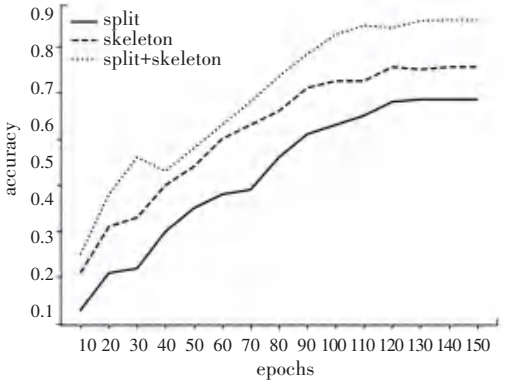

图 9　不同迭代次数下握笔手势分割图、骨架图以及双模态图像的识别率

Fig. 9　Recognition rates of pen－holding gesture segmentation image, skeleton image and bimodal images with different iteration times

3.2.3　第三组实验

比较了本文模型与 AlexNet 模型以及文献[13]中提出的改进 AlexNet 模型的识别精度，用于验证本文改进的 AlexNet 模型在特征提取能力上和识别准确率上的提高，实验结果如图 10 所示。

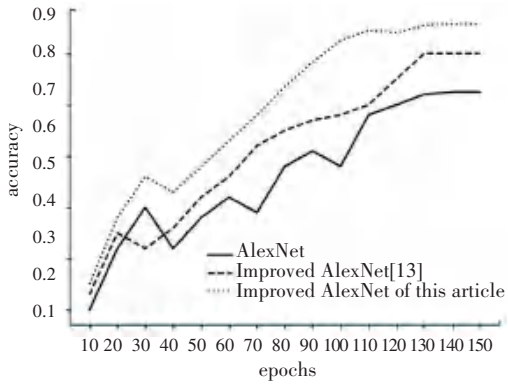

图 10　不同迭代次数下双模态图像输入至 3 种网络的识别率

Fig. 10　Recognition rate of bimodal images input to three networks with different iteration times

由图 10 可知，3 种模型的对比，发现本文模型的识别率高于 AlexNet 以及文献[13]中改进的 AlexNet 模型。由于 AlexNet 网络结构第一个卷积层是 11×11、无注意力机制，使得提取的特征不够丰富有效，且卷积层之后没有添加 BN 层，使得识别率波动大；文献[13]中虽然提出了添加 BN 层以及调参的方法，但是没有添加注意力机制，使得握笔手势图像的重要特征没有被关注。

基于改进 AlexNet 的双模态握笔手势识别的准确率，相比单模态以及其它网络结构有明显提升。同时，不同迭代次数产生的识别效果也有所差别。针对本实验所采用的双模态握笔手势数据集及其预处理操作，迭代次数为 150 的时候所获得的识别效果最好。该实验结果表明，本文提出的模型通过对双模态握笔手势数据集进行特征提取，能够获得相比于单模态数据集更加丰富的特征信息，融合这些特征对握笔手势图像进行分类，能够有效提高卷积神经网络的静态握笔手势识别准确率。

4　结束语

近年来，关于握笔手势的理论研究很多，但是相关人工智能方面的实践却很少。且当下部分中小学生握笔手势不标准，导致坐姿不健康的同时致使眼睛近视、颈椎弯曲以及手指关节增生等问题，对其未来身心健康的发展造成不良的影响。本文初步研究了握笔手势的识别，受到多模态的启发，提出了一种 AlexNet 优化与双模态的握笔手势识别方法，同时自建了握笔手势数据集，实现了 9 种握笔手势识别。未来将进一步扩充握笔手势数据集的同时研究动态的握笔手势识别。

参考文献

[1] 宋一凡，张鹏，刘立波. 基于视觉手势识别的人机交互系统[J]. 计算机科学，2019，46(S2)：570-574.

（下转第 62 页）