Computer Science and Application 计算机科学与应用, 2021, 11(6), 1791-1801
Published Online June 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.116185

Hans汉斯

一种基于改进3D卷积的乒乓球球员动作识别算法

汪语哲[1,2,3]，刘明方[1,2,4]，尹真杰[1,2,4]，段晓东[1,2,4]

[1]大连民族大学大数据应用技术国家民委重点实验室，辽宁 大连
[2]大连民族大学大连市民族文化数字技术重点实验室，辽宁 大连
[3]大连民族大学机电工程学院，辽宁 大连
[4]大连民族大学计算机科学与工程学院，辽宁 大连
Email: duanxd_dmu@163.com

收稿日期：2021年5月26日；录用日期：2021年6月21日；发布日期：2021年6月28日

摘 要

文章研究了基于视频的乒乓球球员动作识别问题。在计算机视觉领域，人体动作识别具有一定挑战性。基于专业乒乓球运动员在乒乓球发球机的接发动作视频，构建了乒乓球击球动作视频数据集，将其分为正手击球、反手击球、正手拉球、反手拉球和非击球动作5类。提出通过人体密集姿态(Dense Pose)处理数据集，将把人体形态从环境中进行提取，随后提出一种改进的C3D卷积网络，用于学习数据集上连续帧的时空特征。结果表明，文章设计的算法对于光线、环境等干扰因素具有较好的鲁棒性，泛化性能好，为基于视频的动作分类识别问题提出了一种可行解决方案。

关键词

乒乓球击球，人体动作识别，3D卷积网络，动作分类，视频跟踪

A Recognition Algorithm of Player Motion of Table Tennis Based on Improved 3D Convolution

Yuzhe Wang[1,2,3], Mingfang Liu[1,2,4], Zhenjie Yin[1,2,4], Xiaodong Duan[1,2,4]

[1]SEAC Key Laboratory of Big Data Applied Technology, Dalian Minzu University, Dalian Liaoning
[2]Dalian Key Lab of Digital Technology for National Culture, Dalian Minzu University, Dalian Liaoning
[3]College of Mechanical and Electronic Engineering, Dalian Minzu University, Dalian Liaoning
[4]College of Computer Science and Engineering, Dalian Minzu University, Dalian Liaoning
Email: duanxd_dmu@163.com

Received: May 26th, 2021; accepted: Jun. 21st, 2021; published: Jun. 28th, 2021

文章引用: 汪语哲, 刘明方, 尹真杰, 段晓东. 一种基于改进 3D 卷积的乒乓球球员动作识别算法[J]. 计算机科学与应用, 2021, 11(6): 1791-1801. DOI: 10.12677/csa.2021.116185

Abstract

Motion recognition of table tennis players based on video is studied in this paper. Recognition of human action is challenging in the field of computer vision. Based on videos of ball strike of professional table tennis players against table tennis ball machine, a data set of ball strike of table tennis players is constructed and divided into 5 catalogs of forehand shots, backhand shots, forehand shots, backhand shots and non-stike action. Dense pose of the human body is used to process the constructed data set and extract human body shape from the environment, and then an improved C3D convolutional network is proposed to learn the spatiotemporal features of continuous frames on the data set. Results show that the algorithm proposed in the article has good robustness to interference factors such as light and environment, and good generalization performance, demonstrating a feasible solution to the problem of video-based action classification and recognition.

Keywords

Striking of Table Tennis, Human Action Recognition, 3D Convolutional Network, Action Classification, Video Tracking

1. 引言

乒乓球被誉为我国的国球，其易于开展，参与性强，普及度高，得到了国内各年龄段人员的普遍喜爱。随着科技的高速发展，计算机技术已成为提升运动竞技水平的重要工具，带有智能功能的比赛记录和分析装置已在训练和比赛中得到了初步应用。王恺凡等人基于卷积神经网络开发了一款乒乓球训练平台，将人脸识别、人体骨骼识别和心率检测功能进行了整合，提升了训练效果[1]。丁朔等人融合语音交互与自然语言处理技术，设计了一款乒乓球智能训练系统，辅助教练员制定训练计划，并对错误的动作给予视频纠正指导[2]。杨波等人将 VR (Virtual Reality，虚拟现实)技术引入高等院校的乒乓球课堂教学，将其整合融入既有课程，提升了教学效果[3]。任云青等人采用 ROS (Robot Operating System，机器人操作系统)、OpenCV (Open Source Computer Vision Library，开源计算机视觉库)等软件，设计并实现了一款智能乒乓球自动捡球机器人，实现了乒乓球智能识别、分拣与运输等功能[4]。

在以乒乓球运动为代表的对抗性球类项目的球员技术动作识别与分析方面，基于机器视觉的人体动作识别技术近来得到了普遍关注。孙于成等使用时空图卷积在其自建的乒乓球骨骼数据集上实现了击球动作的研究[5]。Martin 等提出的双时空卷积神经网络(TSTCNN)在 MediaEval 2020 “运动视频分类：乒乓球的动作分类”比赛中取得了优异的成绩[6]。杨静等基于支持向量机(SVM)和光流分析，提出了一种识别体育视频中羽毛球运动员运动的方法[7]。Nur Azmina Rahmad 等在其自建的羽毛球比赛数据集中对 AlexNet，GoogleNet，VggNet-16 和 VggNet-19 四种卷积神经网络的模型分类性能进行了比较[8]。Piergiovanni 等构建了 MLB-YouTube 棒球比赛运动数据集，并用 3D 卷积网络的方法对数据集细粒度活动进行识别[9]。

然而上述动作识别算法研究均基于公开或非公开球员间比赛的视频数据集，在疫情常态化的新形势下，发球机已成为训练的重要组成部分，发球机发球速度、力量和旋转方式和球员回球有较大差异，而关于球员在乒乓球发球机上的回球动作识别相关研究目前比较罕见。为提升乒乓球发球机的训练质量，课题组邀请专业乒乓球队员进行标注技术动作采集，通过 60 fps 深度摄像机采集专业运动员接发球动作，建立了发

球机接发球标准动作数据集，随后通过密集姿态(Dense Pose)技术去除环境、衣服、肤色等影响因素，最后对 C3D 网络[10]改进并训练数据，完成对乒乓球击球动作的识别，整个项目的实验过程如图 1 所示。

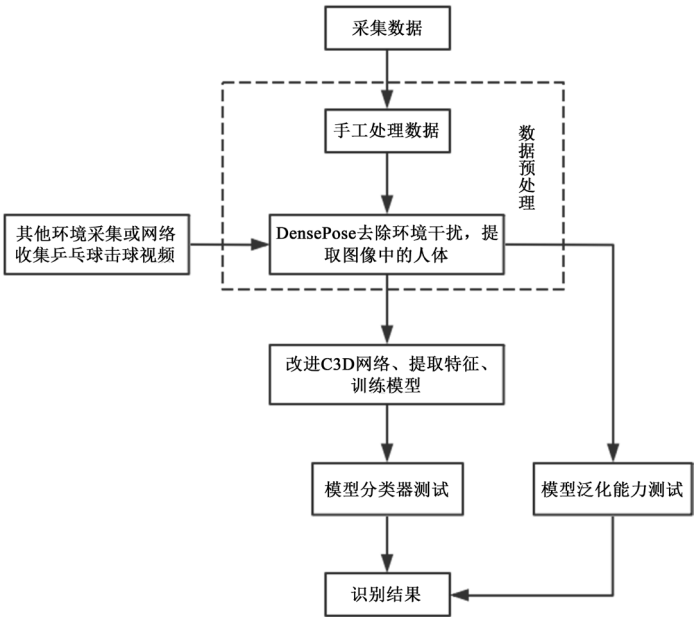


Figure 1. Flow chart of the recognition algorithm of table tennis striking action
图 1. 乒乓球击球动作识别算法流程图

2. 视频数据集构建

2.1. 数据采集环境搭建

选取市场上主流的乒乓球发球机(双鱼 1040)进行接发球训练。该设备在出球口设置了胶皮擦球装置，能够通过胶皮摩擦发出上旋、下旋转和侧旋球；同能够调整发球速度、角度、频率，可以满足接发球技术动作训练要求。

视频采集设备为 ZED 双目深度像机，参数调整为：分辨率 720 像素、帧率 60 fps。为了保证采集到的运动视频不会出现丢帧和跳帧等问题，在数据采集时对连接摄像机的电脑硬件设备进行了调整。整体采集设备硬件如图 2 所示。

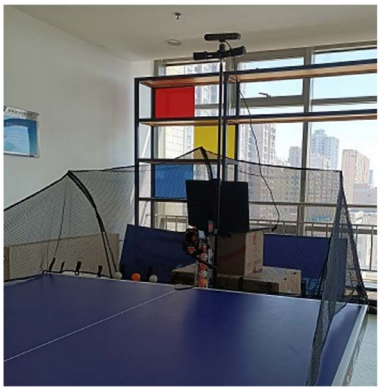


Figure 2. Equipment for data acquisition
图 2. 数据采集设备

2.2. 数据集搭建

目前乒乓球接发球主流的技术动作可分为正反手拉球、正反手攻球、正反手搓球、正反手削球等几类，目前这些技术动作尚无准确的动作集标准，然而统计分析表明，多数高水平运动员在乒乓球发球机上的接发球动作有较大相似性，因此选择邀请高水乒乓球球员动员录制接发球视频，将他们的击球动作作为标准数据集。

香港中文大学研究者于 2020 年构建了一个大规模体操运动人体动作数据集：FineGym [11]，数据集通过层级化标注对动作的细粒度进行了区分，最细划分到一个具体的体操动作，例如平衡木体操下马动作中的团身前空翻动作。文章根据此数据所划分的细粒度动作，构建了乒乓球击球动作数据集。

乒乓球击球动握拍方式大致可以分为横拍和直拍两种，通过分析视频中的连续帧发现，直拍中有些击球动作轨迹相似，例如正手攻球和正手拉球在即将击球的瞬间的连续 2~3 帧动作相似，反手攻球和反手拉球在准备接球和击球的瞬间都有 1~3 帧的动作相似，这也是击球动作中区分的难点，击球动作相似和动作差异如图 3 所示。因此，在采集数据和做动作区分时，选择了对正手攻球、正手拉球、反手攻球、反手拉球的动作进行采集。采集的数据集中，击球动作之外的动作将其称为其他动作，对 4 类击球动作和 1 类其他动作进行识别，以便于模型更好区分击球动作。

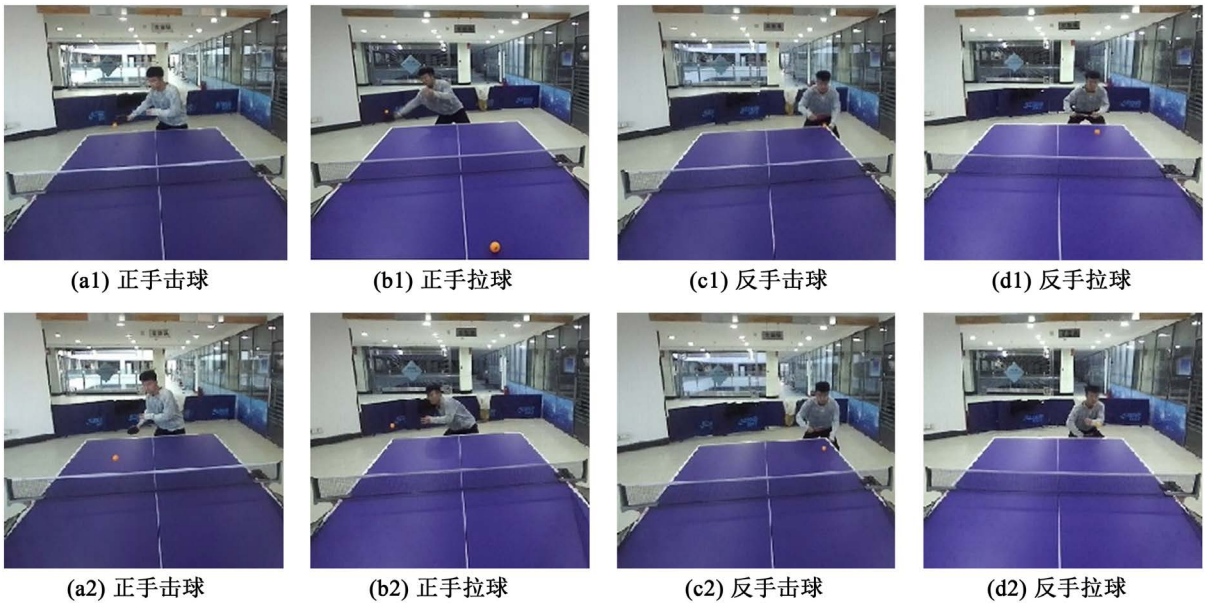


Figure 3. The similar action in the stroke action: (a1) and (b1) are the frames of the similar action in the forehand stroke and forehand pull; (a2) and (b2) are frames of the difference between the forehand stroke and the forehand pull. (c1) and (d1) are frames of similar movements in the backhand stroke and backhand pull

图 3. 击球动作中的相似动作：(a1)和(b1)为正手击球和正手拉球过程中相似动作的某一帧；(a2)和(b2)为正手击球和正手拉球过程中动作区别的某一帧；(c1)和(d1)为反手击球和反手拉球过程中的相似动作的某一帧

3. 数据集处理

3.1. 数据预处理

对所采集的乒乓球击球视频进行处理。由于在乒乓球运动中，击球是一个快速动作，直接给视频数据设置标签不仅比较困难，而且标签所在的时间段也存在其他动作的干扰。

通过将乒乓球击球视频的每一帧提出来，经过研究对比发现，每种击球动基本在 16~24 帧内完成

击球，因此将完全包含一个击球动作的 16~24 帧的连续帧进行提取，形成一个击球动作的时空特征数据；非击球动作一般在 48 个连续帧左右之内，将提取非击球动作的 48 个连续帧，形成一个非击球动作的时空特征数据。经过手工提取数据后，数据集包括了 824 个动作，将近 20,000 帧图片，5 类动作每类包含 150 左右个动作连续帧的时空特征数据。表 1 完整的显示了动作类别和动作数量。为了便于训练、Dense Pose 处理数据和设置动作标签，设置了五个文件夹，其名字作为应五个动作类的标签，既 Facade Attack (正手攻球)、Facade Pull (正手拉球)、Back Attack (反手攻球)、Facade Attack (反手拉球)和 Other Action (非击球动作)五类运动标签。每个标签文件夹中的每一个子文件夹都代表一个动作，每个动作由连续帧组成。

Table 1. Data set of action category and number of actions
表 1. 数据集动作类别和动作数量

击球动作	动作样本数量
正手击球	179
正手拉球	122
反手击球	206
反手拉球	158
其他动作	159

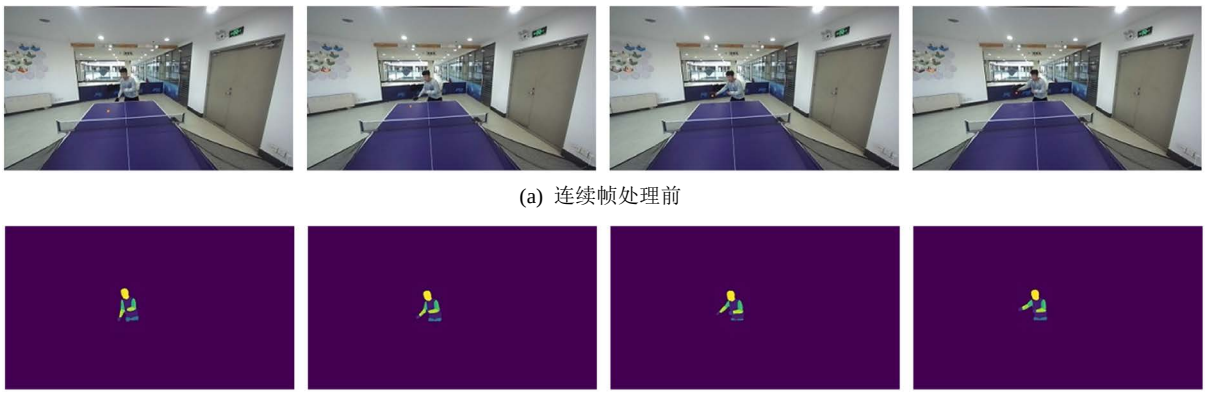
3.2. Dense Pose 处理数据集

2018 年 Facebook [12]和 Inria France [13]的研究者分别在 ECCV 会议和 CVPR 会议发表了篇有关于 Dense Pose 的文章，介绍了他们提出的 Dense Pose 系统[14]。在 Dense Pose 系统中的密集姿态估计(Dense Pose Estimation)功能[11]可以将 2D 人体映射到 3D 的人体表面，在图 4 中展示了 2D 到 3D 的转换。这种 2D 映射到 3D 转换的过程并不会改变图像的大小，所映射的 3D 人体表面在图像中分成了 24 份，每一份的中的颜色值存在偏差，并且此功能可去除图像中的背景干扰，如图 4 所示。

(a) 处理前　　(b)处理后

Figure 4. Pose estimation of image processing using Dense Pose system
图 4. Dense Pose 系统密集姿态估计处理图像

通过 Dense Pose 系统中的密集姿态估计功能，对手工处理后的数据集中动作帧图像进行处理，把 2D 人体映射到 3D 模型，经过此处理图片中的人体只有形态和动作，模型的泛化效果不会被人体的肤色和着装所影响。经过 Dense Pose 处理过的图像，动作识别中环境干扰因素都会被去除，模型泛化效果就不会受环境等因素所影响。图 5 介绍了未处理的数据对比经过处理后的击球部分连续动作帧。



(a) 连续帧处理前

(b) 连续帧处理后

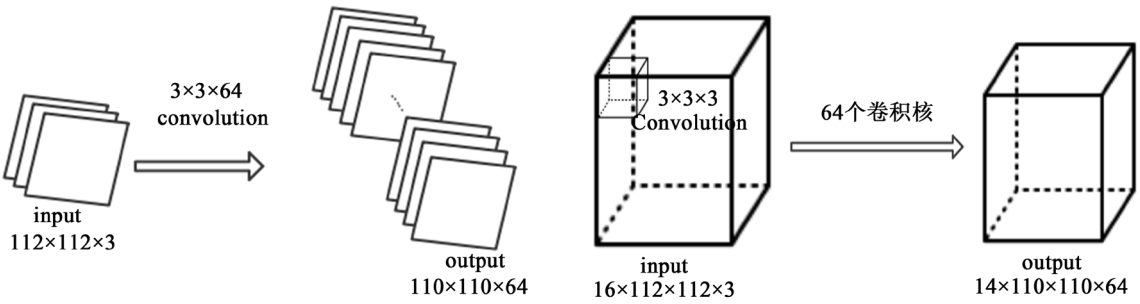
Figure 5. Comparison of unprocessed data with continuous action frame of striking of table tennis processed by Dense Pose
图 5. 未处理的数据对比 Dense Pose 处理后的击球部分连续动作帧

4. 网络模型搭建

4.1. 3D 卷积网络

3D 卷积也是从 2D 卷积的基础上发展而来的，基于 2D 卷积核的卷积网络[15]可用于学习单通道或者三通道的图像的空间特征，图 6 中的图 6(a)是 2D 卷积应用在三通道的图像卷积的结果，输入为 w × h × c 三个维度，表示图片的宽、高和通道数。2D 卷积应用在学习单通道图像的特征时，除了输入维度少一个通道维度外，整个卷积过程与图 6(a)三通道卷积过程相同。

相比于 2D 卷积核来说 3D 卷积核[10]是多了一个维度去学习数据中的时间特征，如图 6(b)输入为 t × w × h × c 四个维度，其中 t 代表连续帧数目，t 这一维度的特征就是数据时间维度上的特征。对于连续的单通道图像来说，输入的维度是 t × w × h 三个维度，整个卷积过程和图 6(b)三通道图像卷积过程相同。



(a) 2D 卷积过程　　(b) 3D 卷积过程

Figure 6. Difference between 2D convolution process and 3D convolution process
图 6. 2D 卷积过程与 3D 卷积过程的区别

4.2. C3D 网络结构改进

C3D 网络是 Du Tan (杜兰特)等人[10]于 2015 年提出的一个用于学习和识别视频中信息的通用网络，其通过 3D 卷积网络去学习视频中的时空特征，学习所得到的模型可用于场景分类、动作识别等领域。C3D 网络的训练所使用的数据集为 UCF101 数据集[16]，数据集由 101 个人类动作类别的 13,320 个视频组成，视频图像未经过处理，由于数据集种类多，视频图像中的特征复杂，因此 C3D 网络整体网络层次

设计比较深，包含为 8 个卷积层、5 个池化层、2 个全连接层，整个网络的参数为 6000 万左右。在项目组自建的数据集中，数据经过处理后没有了环境、衣服、肤色等特征的影响，图像中只有人体形态，图像中的特征相对简单，为了防止出现过拟合现象，因此将 C3D 网络的整体结构缩减为 5 个卷积层、5 个池化层、2 个全连接层，并对特征图个数和神经元个数进行了缩减，最终网络参数降为 500 万左右。整个网络结构如图 7。

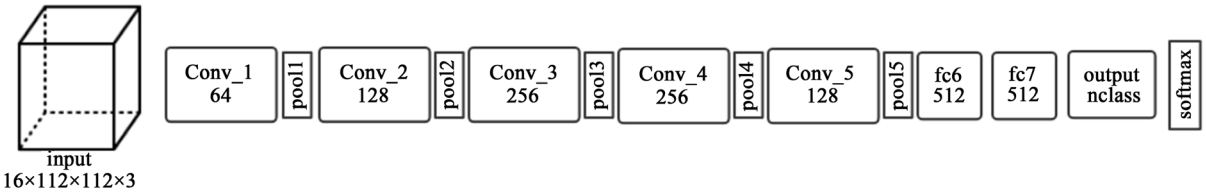


Figure 7. Structure diagram of improved C3D network
图 7. 改进后的 C3D 网络结构图

输入层(Input)：输入的数据为 16 个连续的视频帧，每一帧的宽高为 112 × 112，通道数为 3，即输入维度为 16 × 112 × 112 × 3；

卷积层(Convolution)：5 个卷积核的每一个卷积核维度为 3 × 3 × 3，步长为 1，并在卷积过程让其自动填充 padding，保证卷积过程中输入和输出尺寸相同，在每个卷积过程结束后，使用 ReLU 激活函数。卷积计算公式如式(1)，其中 W 为帧的宽度，H 为帧的高度，T 为连续帧的时间维度，p 为边界填充值，s 为步长，为卷积核维度。

$$\begin{cases} W_{out} = \dfrac{W_{in} + 2p - C_W}{s} + 1 \\ H_{out} = \dfrac{H_{in} + 2p - C_H}{s} + 1 \\ T_{out} = \dfrac{T_{in} + 2p - C_T}{s} + 1 \end{cases} \tag{1}$$

池化层(Polling)：池化层使用的事最大值池化，为了在初始卷积池化阶段保留时间上的特征，保留了原网络的池化结构，在第一次池化时维度为 1 × 2 × 2，步长为 1 × 2 × 2，其他池化层的维度都为 2 × 2 × 2，步长为 2 [6]；

全连接层(FC)：全连接层的神经元个数根据最后一次池化后特征图的大小设置为 512 个。经过前面卷积提取特征后，全连接层最终会将这些特征映射到标签空间，起到一个分类器的作用。

输出层(output)：根据数据集中 5 个类别的动作，网络最后的输出是判断的这五个类别的概率值。

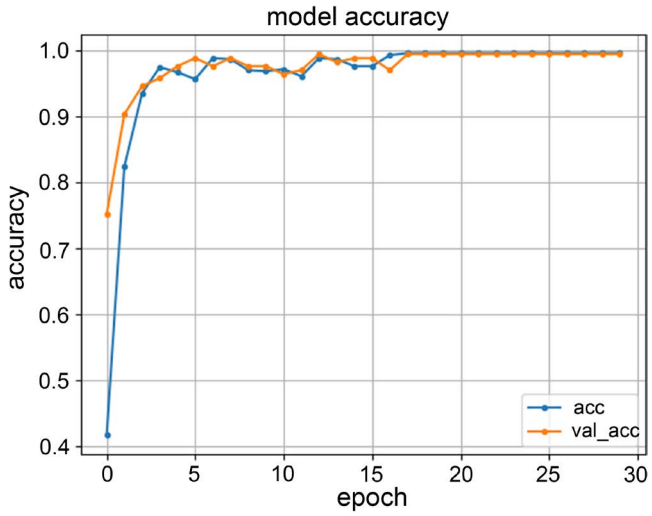
5. 正文实验结果与分析

基于自建数据集训练乒乓球击球动作识别模型实验硬件环境为：处理器 Intel(R) Core(TM) i7-7800X CPU @ 3.50 GHz 3.50 GHz 和 NVIDIA GeForce RTX 2080Ti GPU；模型泛化性能测试实验环境：硬件环境 Intel(R) Xeon(R) E-2224 CPU @ 3.40 GHz 3.41 GHz 和 NVIDIA GeForce GTX 1650 GPU，DensePose 系统不能再 Windows 系统上安装，所以系统环境为 Ubuntu 操作系统。

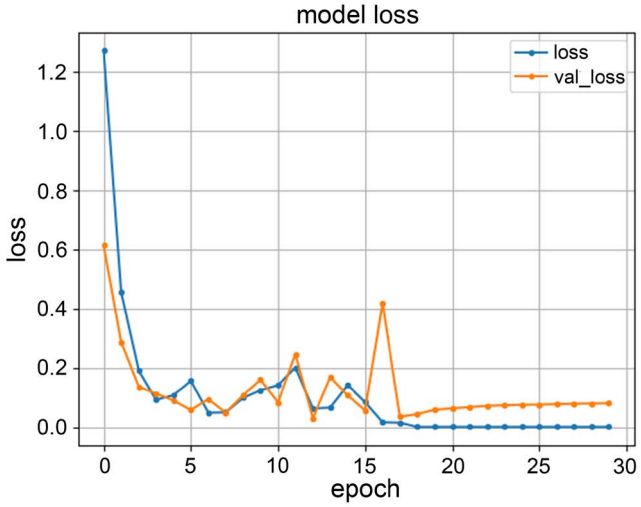
5.1. 模型训练实验结果分析

用于训练模型的训练集和验证集均来自于自建数据集，训练集和测试接的划分别为数据集的 80%和 20%。通过 30 次迭代训练，通过图 8(a)可以看出在第 18 次迭代后训练集精准度稳定在 99.5%，验证集精

准度稳定在 99.3%。通过图 8(b)可以看出在第 18 次迭代后训练集损失函数值趋近于 0，验证集损失函数值稳定在 0.09 以内。



(a) 训练集和验证集训练结果准确率曲线

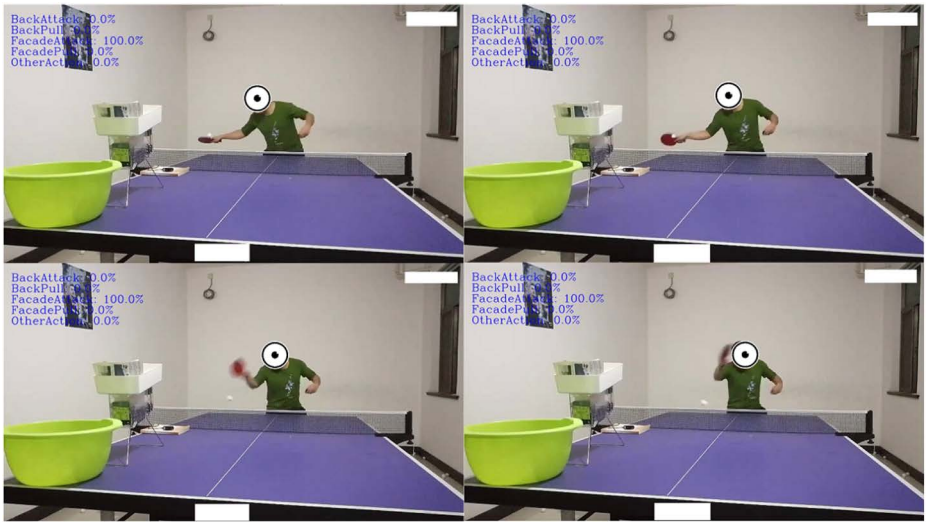


(b) 训练集和验证集训练结果损失函数值曲线

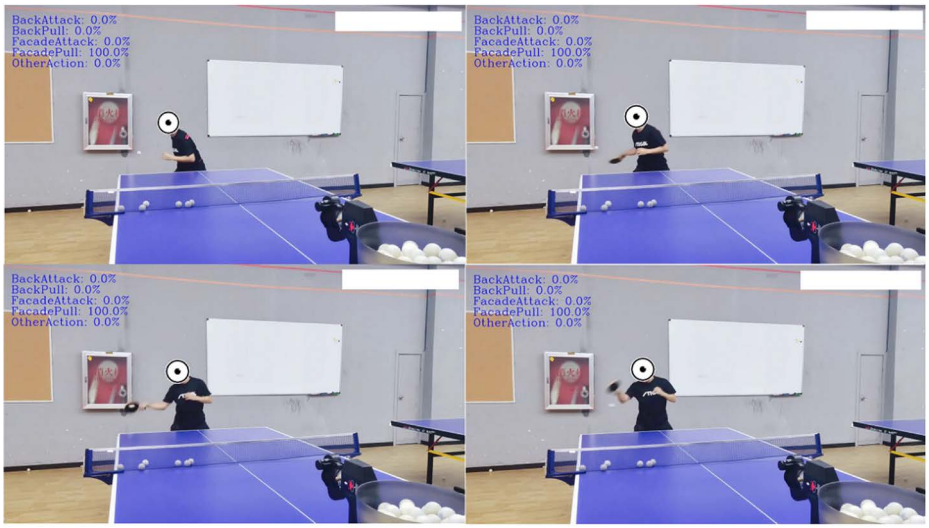
Figure 8. Accuracy and loss function value curves of the training results of the training set and the validation set
图 8. 训练集和验证集训练结果的准确率和损失函数值曲线

5.2. 模型泛化效果预测

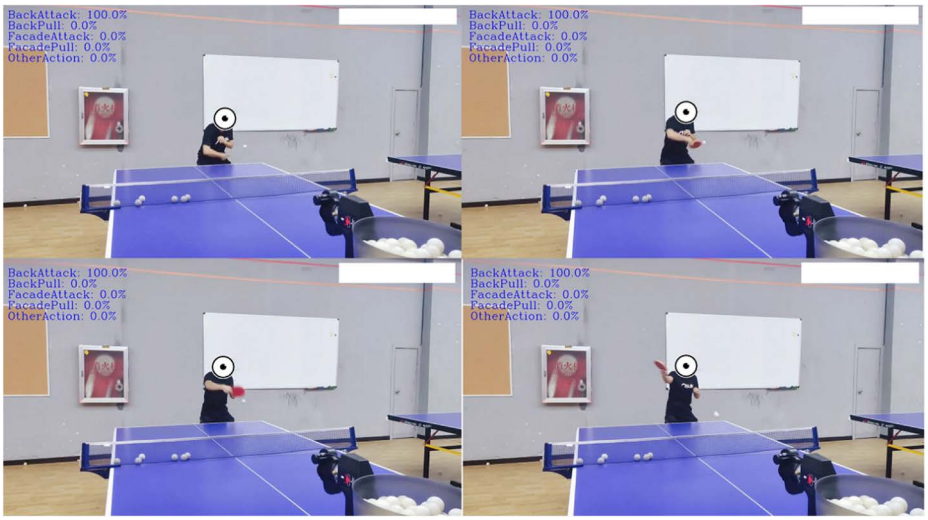
乒乓球击球动作识别模型泛化效果的好坏决定着这个模型是否具有实际应用价值，通过从网络上找到拍摄角度与课题组自建数据集中拍摄角度相似的乒乓球击球视频，整个视频经过 Dense Pose 系统处理后放入到模型中测试泛化效果。图 9 中的视频来源于视频网站与训练集中的数据不存在交集，并且图中为连续动作帧中提取出来图像。图 9 中，图 9(a)预测结果为正手攻球，图 9(b)预测结果为正手拉球，图 9(c)预测为反手攻球，图 9(d)预测为反手拉球，预测结果均与球员实际击球动作类型一致。通过泛化性能测试可以看出本文所提出的方法训练出来的动作识别模型泛化能力较强，有一定的实际应用价值。



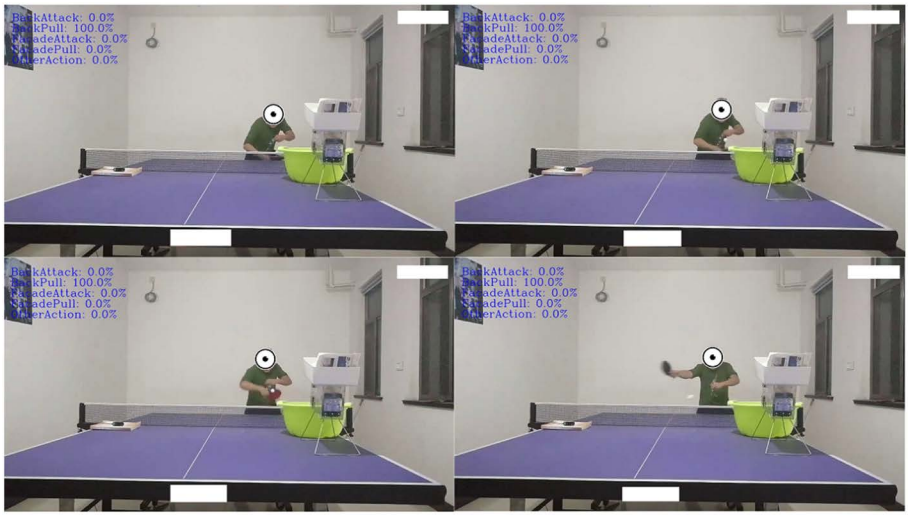
(a) 预测为正手攻球



(b) 预测为正手拉球



(c) 预测为反手攻球



(d) 预测为反手拉球

Figure 9. Generalization performance test of recognition model for table tennis striking action
图 9. 乒乓球击球动作识别模型泛化性能测试

6. 结语

基于专业运动员在乒乓球发球机上的接发球视频，自建了包含五分类的乒乓球击球动作识别视频数据集，通过动作划分手工提取了连续击球动作视频帧，采用 Dense Pose 系统中的密集姿态估计功能对手工提取的视频帧进行了处理，最终通过改进 C3D 网络训练出了乒乓球击球动作识别模型。实验表明：模型训练时在第 17 次迭代之后训练精准度稳定在 99.5%，验证集精准度稳定在 99.3%，并且模型通过了泛化性能测试，证明文章提出的方法可以去除环境、衣服等因素的干扰，训练的乒乓球击球动作识别模型具有实际应用的价值。

基金项目

大连市科技创新基金项目“面向足球青训的技战术分析算法及配套可穿戴设备研发”(项目编号：2020JJ26GX038)；大连民族大学学科团队项目“基于机器学习的乒乓球接发球动作识别与水平评估算法研究”。

参考文献

[1] 王恺凡. 基于人脸识别的乒乓球智能训练平台设计[D]: [硕士学位论文]. 南京: 南京邮电大学, 2020.

[2] 丁朔. 基于智能语音交互的乒乓球训练系统的设计与实现[D]: [硕士学位论文]. 南京: 南京邮电大学, 2020.

[3] 杨波. 虚拟现实技术应用于高校乒乓球教学中的实证研究[D]: [硕士学位论文]. 兰州: 西北师范大学, 2020.

[4] 任云青. 智能乒乓球自动捡球机器人的设计与实现[D]: [硕士学位论文]. 南京: 南京邮电大学, 2020.

[5] 孙于成. 基于时空图卷积的乒乓球基础技术动作识别[D]: [硕士学位论文].安庆: 安庆师范大学, 2020.

[6] Martin, P.-E., Benois-Pineau, J., Péteri, R. and Morlier, J. (2020) 3D Attention Mechanism for Fine-Grained classification of table tennis strokes using a Twin Spatio-Temporal Convolutional Neural Networks. 25*th International Conference on Pattern Recognition*, Milano, January 2021. arXiv preprint arXiv:2012.05342.

[7] 杨静. 体育视频中羽毛球运动员的动作识别[J]. 自动化技术与应用, 2018, 37(10): 120-124.

[8] binti Rahmad, N.A., binti Sufri, N.A. J, bin As'ari, M.A., *et al.* (2019) Recognition of Badminton Action Using Convolutional Neural Network. *Indonesian Journal of Electrical Engineering and Informatics* (*IJEEI*), **7**, 750-756. https://doi.org/10.11591/ijeei.v7i4.968

[9] Piergiovanni, A.J. and Ryoo, M.S. (2018) Fine-Grained Activity Recognition in Baseball Videos. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Salt Lake City, 18-22 June 2018, 1740-1748. https://doi.org/10.1109/CVPRW.2018.00226

[10] Tran, D., Bourdev, L., Fergus, R., *et al.* (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 *IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[11] Shao, D., Zhao, Y., Dai, B. and Liu, D. (2020) FineGym: A Hierarchical Video Dataset for Fine-Grained Action Understanding. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 13-19 June 2020, Seattle, 2616-2625. https://doi.org/10.1109/CVPR42600.2020.00269

[12] Güler, R.A., Neverova, N. and Kokkinos, I. (2018) Densepose: Dense Human Pose Estimation in the Wild. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7297-7306. https://doi.org/10.1109/CVPR.2018.00762

[13] Neverova, N., Guler, R.A. and Kokkinos, I. (2018) Dense Pose Transfer. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Proceedings of the European Conference on Computer Vision* (*ECCV*), Springer, Cham, 128-143. https://doi.org/10.1007/978-3-030-01219-9_8

[14] 机器之心 Pro.Facebook 实时人体姿态估计: Dense Pose 及其应用展望[EB/OL]. https://baijiahao.baidu.com/s?id=1625055353488715502&wfr=spider&for=pc, 2019-02-01.

[15] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[16] Soomro, K., Zamir, A.R. and Shah, M. (2012) UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *CoRR*, **1212**, 0402.