arXiv:2302.02755v1 [cs.CV] 6 Feb 2023

使用双流卷积网络结合 RGB 和姿态信息进行细粒度动作检测

LeonardHacker[1], *, FinnBartels[1], * 和Pierre-Etienne Martin[2], †

[1]计算机科学研究所，莱比锡大学，德国[2]*CCP*系，马克斯·普朗克进化人类学研究所，*D-04103* 莱比锡，德国

摘要

作为 MediaEval 2022 体育任务的参与者，我们提出了一种用于乒乓球击球动作分类和检测的双流网络方法。每个流都是由使用注意力机制的三维卷积神经网络（CNN）模块串联而成。每个流处理不同的四维输入。我们的方法利用原始 RGB数据和通过 MMPose工具箱计算得到的姿态信息。通过将姿态应用于黑色背景或其来源的原始 RGB帧上，将姿态信息作为图像处理。最佳性能通过向一个流输入原始 RGB数据，向另一个流输入 Pose + RGB（PRGB）信息，并对特征应用后期融合获得。在提供的 TTStroke-21 数据集上对这些方法进行了评估。我们可以报告，击球动作分类性能有所提升，准确率达到 87.3%；而检测性能虽未超过基线，但仍达到 0.349 的交并比和 0.110 的平均精度均值。

1. 引言

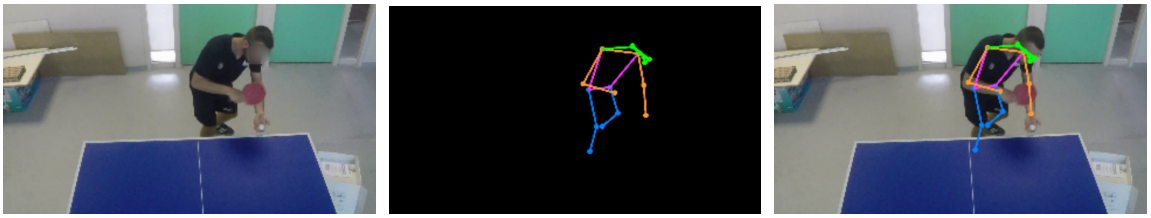
尽管视频中的粗粒度动作检测（例如，识别正在进行的运动项目）已取得长足进展，但由于类间差异较小，细粒度动作检测本质上更具挑战性 [1, 2]。这项基准任务旨在提供可行的工具，以便分析运动员表现 [2]。乒乓球的细粒度视频检测面临许多有趣的挑战，例如球轨迹预测 [3] 或实时比分与比赛分析 [4]。

在图像识别领域，由卷积层、ReLU层和最大池化层组成的卷积神经网络是一种常见做法 [5]。竞赛提供的基线模型源自这一方法 [6, 7]。虽然视频数据增加了一个维度（时间或帧编号），但有几种方法会调整普通卷积神经网络，以确定一系列帧中的运动或变化 [1]。由 Simonyan 和 Zisserman [8] 首先提出的一种流行方法是双流神经网络。他们实现了一种由空间卷积神经网络和时间卷积神经网络组成的架构。上述空间流在每个时刻表示单独的一帧，而时间
该流由多帧光流（OF）构成，用于识别多个 RGB 帧中的运动动量。因此，该模型还能获得互补的

MediaEval'22：多媒体评测研讨会，*2023*年*1*月*13*日至*15*日，挪威卑尔根及线上 *这些作者贡献相同。

†通讯作者。 pierre_etienne_martin@eva.mpg.de (P. Martin)
www.eva.mpg.de/comparative-cultural-psychology/staff/pierre-etienne-martin (P. Martin) 0000-0002-9593-4580 (P. Martin)

© 2022 年本文作者版权所有。允许根据知识共享署名 4.0 国际许可协议（CC BY 4.0）使用。CEUR研讨会论文集（CEUR-WS.org）



(a) RGB (b) 姿态 (c) PRGB

图1： TTStroke-21 视频中的 RGB、姿态和 PRGB 帧。生成的帧用作网络分支的输入。

信息，这些信息由第二条流提供。因此，可以轻松移除视频中没有运动的区域，以减少噪声 [1]。基于此成功实现的模型是膨胀三维卷积神经网络（I3D）模型 [9],，它将多张图像输入三维卷积神经网络，而非每次只输入一张图像。 Feichtenhofer 等人 [10] 提出了基于双流架构的 SlowFast 网络。第一条流以较低帧率处理帧，第二条流则以较高帧率处理帧，从而兼顾语义信息和运动信息。

融合步骤是多流网络的重要组成部分。在这些模型中，根据融合执行的位置，有多种不同的方法。后期融合是最简单且最高效的方法之一 [11, 1]。融合可以在全连接 ReLU 层之后、最终的 Softmax 函数之前执行 [12, 13]。早期融合通常在网络的较早阶段执行，这意味着第二条流的信息被推送到第一条流中 [10, 14]。以下章节将介绍双流架构、击球动作分类和击球动作检测的结果，以及对初步因素和方法的讨论。

2. 方法

如前所述，双流架构是最先进的方法之一。在本项工作中，我们将 MediaEval 提供的基线扩展为一个双流网络，使用原始 RGB 图像和姿态信息。基线本身是一个带有注意力机制的单流三维卷积神经网络。近期论文表明，与光流相比，利用姿态信息能够在细粒度动作检测中取得更好的结果 [15, 16, 17]。对于由人执行的动作，姿态包含大量重要信息。我们通过将姿态信息绘制在 RGB 图像上，将其叠加起来，并将结果输入第二个流中。我们的方法实现可在线获取[1]。

2.1. 姿态估计

姿态信息是每一帧中描绘出的人体姿态，如图 1所示。由于光流已得到充分研究，且使用姿态信息能取得良好效果，因此本研究着重探讨如何在双流卷积神经网络框架中使用姿态信息。姿态信息通过 OpenMMLab 的 MMPose [18] 工具箱提取。该工具箱是一个用于细致视频理解的开源软件包。系统会分析输入视频的每一帧。首先，使用人体检测器在

[1]https://github.com/fidsinn/SportTaskME22

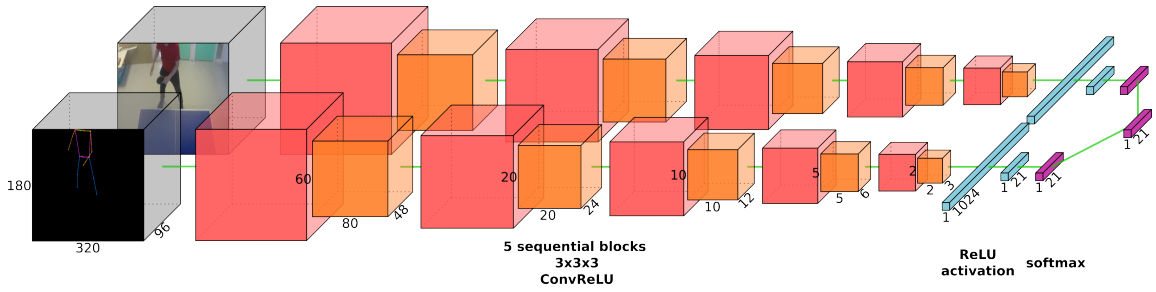


图 2：双流姿态卷积神经网络：上方分支使用 RGB图像，下方分支使用黑色背景上的姿态信息。每个分支由五个连续的卷积层组成，使用 ReLU激活函数，滤波器数量逐层递增（32、64、128、256、512）；每层后均接一个注意力块（均以红色表示）和一个池化层（以橙色表示）。前两个池化层的尺寸为 $(4 \times 3 \times 2)$，其余池化层的尺寸为 $(2 \times 2 \times 2)$。

每个人周围绘制边界框，然后使用姿态估计器从边界框中提取姿态。人体检测采用更快的基于区域的卷积神经网络 [19]，姿态估计则采用深度高分辨率表征学习 [20]。关键点提取使用自顶向下分类器，因为其表现优于自底向上分类器 [15, 21]。这两种模型均由 OpenMMLab 使用 COCO数据集 [21] 进行预训练。随后，用不同颜色的线段连接各个关键点，以表示身体部位，并将其叠加在图像上。由于不同身体部位对笔画的影响各不相同，我们假设模型能够根据各身体部位的颜色将其区分开来。本研究考察了两种利用姿态信息的方法：Pose 和 PRGB。Pose 变体将计算得到的姿态显示在黑色背景上，而 PRGB 则以原始 RGB帧为背景。通过比较 Pose、PRGB 和 RGB 的表现，可以确定网络对姿态信息的利用效果。

2.2. 架构

我们的模型是双流架构的一种变体 [22]，也是对所提供基线的扩展 [7],，该基线同样使用 3D 卷积块和注意力机制。如图 2 所示，双流姿态卷积神经网络（TSPCNN）由两个相同的流组成，每个流包含五个卷积层和池化层，滤波器数量逐层增加，最终连接到一个带 ReLU激活的线性层。该层的输出馈入第二个线性层，再经 Softmax 函数转换为一个 21 维概率向量，用于分类（21 种不同的笔画类型，包括非笔画类别）；或者转换为一个二维向量，用于检测（笔画类别和非笔画类别）。随后，将两个分支的输出相加，并通过最后一个 Softmax 函数处理，以获得用于分类和检测的概率输出。第一个 Softmax 函数会在融合前对每个独立流的输出进行归一化，以尽量减少梯度消失。

2.3. 融合

文献表明，将早期融合与后期融合结合使用，能够显著提升双流模型的性能 [15]。在 TSPCNN 中加入多种融合方法，性能提升有限。采用后期融合方法（即

表 1基线、单流方法和双流方法在分类准确率（%）和检测指标（准确率、交并比、平均精度均值（%））上的结果。

模型	分类			检测			
	训练	验证	Test	训练	验证	测试交并比	测试平均精度均值
基线	-	0.813	0.864[2]	-	-	**0.515**[5] (0.365[3])	**0.131**[5] (0.118[4])
Pose	0.995	**0.878**	0.847[1]	0.862	0.591	0.205[1]	0.046[1]
PRGB	0.978	0.813	0.864[1]	0.980	0.834	0.165[1]	0.036[1]
RGB 与姿态	1	0.830	0.872[1]	0.987	0.820	0.331[1]	0.100[1]
RGB 与 PRGB	0.998	0.848	**0.873**[1]	0.990	**0.840**	0.349[1]	0.110[1]

决策方法：[1]无窗口，[2] 高斯，[3] 均值，[4] 投票，[5] 投票（滑动窗口）

在最后一层之前进行融合）取得了最佳性能。融合方式包括：i）加权融合，所得特征等于两个融合特征的加权和；ii）求和融合，所得特征为两个特征之和；iii）拼接融合，所得特征由两个特征拼接而成。求和融合的表现最佳，因此本文后续仅报告其结果。

2.4. 训练

所有实验均在莱比锡大学提供的 Tesla V100 GPU 上进行。检测任务每轮训练耗时 7 至 8 分钟，分类任务每轮训练耗时 1 至 2 分钟，共训练 2000 轮，学习率为 0.0001，动量为 0.5。

3. 结果与讨论

我们使用 MediaEval Sport 任务组织者提供的 TTStroke-21 数据集 [23] 评估了我们的方法 [2]。我们将结果与提供的基线 [7]进行比较。为了将我们的方法与基线进行比较，我们在分类任务中使用准确率、在检测任务中使用交并比和平均精度均值来评估各次运行结果。此外，我们还列出了每个模型的训练准确率和验证准确率，见表 1。测试结果依据多种决策方法中表现最佳的方法选出 [7]。基线及其他单流方法在分类任务中已取得相当令人满意的结果。基础 RGB 与 PRGB 结合的双流方法在测试中准确率最高。与单流方法相比，双流方法使分类结果略有提升，最高提升 0.009。与分类任务中的提升相反，在检测任务中，采用姿态流的双流方法，其表现低于单流基线。检测表现不佳可能是因为姿态数据缺少球和球拍信息。没有球拍的手臂运动也可能与击球动作相似，从而导致模型混淆。因此，我们可以认为，姿态信息并未提升击球动作检测的表现。

我们已经证明，姿态信息适用于细粒度动作分类，但似乎无法捕捉用于检测的判别特征。虽然 TSPCNN 在分类任务中的表现优于基线，但我们未能提升检测性能。分类准确率略有提升的一个可能原因是训练数据有限，尤其是因为

有些类别在训练集中只有少量带标签视频。

4. 总结与展望

由于可穿戴系统可能造成干扰、设置繁琐，而且并不普及，因此，运动员和教练非常关注基于视频的细粒度动作检测，以便对比赛中的不同动作进行分类并提高训练效率 [2]。我们基于最先进的研究构建了 TSPCNN。我们的主要贡献是在双流网络中使用叠加了人体姿态的 RGB数据。我们的方法在分类准确率方面略优于基线，但击球动作检测的表现不佳。为了进一步改进 TSPCNN，还需要针对不同质量的姿态数据开展更多实验。此外，还可以评估姿态数据的不同表示方式，例如使用更粗的线条来突出姿态。还应在不同的数据集上验证该方法，例如 FineGym 数据集 [24] ，因为该数据集的类别间差异也较小，但训练数据中的类别分布更加均衡。

参考文献

[1] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, M. Li, 深度视频动作识别综合研究，arXiv预印本 arXiv:2012.06567（2020）。[2] P. Martin, J. Calandre, B. Mansencal, J. Benois-Pineau, R. Péteri, L. Mascarilla, J. Morlier, Sport task：MediaEval 2022中基于视频的乒乓球击球动作细粒度检测与分类，载于：MediaEval，CEUR研讨会论文集，CEUR-WS.org，2022。[3] H. Lin, Z. Yu, Y. Huang，乒乓球机器人的球追踪与轨迹预测，Sensors 20（2020）333。[4] R. Voeikov, N. Falaleev, R. Baikulov，TTNet：乒乓球时空视频实时分析，载于：IEEE/CVF CVPR研讨会，2020，第884–885页。[5] S. Albawi, T. A. Mohammed, S. Al-Zawi，卷积神经网络理解，载于：2017年国际工程与技术会议（ICET），2017，第1–6页。[6] P. Martin，MediaEval 2021基准体育视频任务的时空卷积神经网络基线方法，载于：MediaEval，CEUR研讨会论文集，CEUR-WS.org，2021。[7] P. Martin，采用注意力机制和三维卷积神经网络的MediaEval 2022基准体育任务基线方法，载于：MediaEval，CEUR研讨会论文集，CEUR-WS.org，2022。[8] K. Simonyan, A. Zisserman，基于双流卷积网络的视频动作识别，载于：神经信息处理系统进展27（2014）。[9] J. Carreira, A. Zisserman，动作识别何去何从？一种新模型与Kinetics数据集，载于：IEEE CVPR，2017，第6299–6308页。[10] C. Feichtenhofer, H. Fan, J. Malik, K. He，SlowFast网络用于视频识别，载于：IEEE/CVF国际计算机视觉会议，2019，第6202–6211页。[11] P. Martin, J. Benois-Pineau, B. Mansencal, R. Péteri, J. Morlier，MediaEval 2020中采用三流时空卷积神经网络的乒乓球击球动作分类，载于：MediaEval 2020研讨会论文集，线上，2020年12月14–15日，2020。[12] P.-E. Martin，基于时空卷积神经网络的视频细粒度动作检测与分类：在乒乓球中的应用，博士论文，波尔多大学；拉罗谢尔大学，2020。网址：https://hal.archives-ouvertes.fr/tel-03099907。[13] P. Martin, J. Benois-Pineau, R. Péteri, J. Morlier，采用孪生时空卷积神经网络的体育动作识别：在乒乓球中的应用，载于：CBMI，IEEE，2018，第1–6页。

[14] C. Feichtenhofer, A. Pinz, A. Zisserman，面向视频动作识别的卷积双流网络融合，载于：IEEE CVPR，2016，第1933–1941页。[15] H. Duan, Y. Zhao, K. Chen, D. Shao, D. Lin, B. Dai，重新审视基于骨架的动作识别，arXiv预印本 arXiv:2104.13586（2021）。[16] P. Martin, J. Benois-Pineau, R. Péteri, J. Morlier，采用三维卷积网络进行细粒度动作分类的运动估计方法优选，载于：ICIP，IEEE，2019，第554–558页。[17] S. Sato, M. Aono，利用人体姿态估计模型进行乒乓球击球动作分类，载于：MediaEval，CEUR研讨会论文集第2882卷*CEUR*研讨会论文集，CEUR-WS.org，2020。[18]M. Contributors，OpenMMLab姿态估计工具箱与基准，https://github.com/open-mmlab/mmpose，2020。[19] S. Ren, K. He, R. Girshick, J. Sun，Faster R-CNN：借助区域建议网络实现实时目标检测，载于：神经信息处理系统进展28（2015）。[20] K. Sun, B. Xiao, D. Liu, J. Wang，面向人体姿态估计的深度高分辨率表征学习，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2019，第5693–5703页。[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick，Microsoft COCO：情境中的常见物体，载于：ECCV，Springer，2014，第740–755页。[22] P. Martin, J. Benois-Pineau, R. Péteri, J. Morlier，双路时空卷积神经网络中的三维注意力机制：在乒乓球比赛视频动作分类中的应用，载于：ICPR，IEEE计算机学会，2021。[23] P. Martin, J. Benois-Pineau, R. Péteri, J. Morlier，采用双路时空卷积神经网络的细粒度体育动作识别，Multim. Tools Appl. 79（2020）20429–20447。[24] D. Shao, Y. Zhao, B. Dai, D. Lin, FineGym：用于细粒度动作理解的分层视频数据集，载于：CVPR，IEEE，2020，第2613–2622页。