基于改进的双流三维卷积神经网络的人体行为识别

罗 伟[1] 陈 俊[2]
LUO Wei CHEN Jun

摘 要 为提升在相机运动、场景变化等复杂背景视频中人体行为的识别准确率，并获得视频的全局上下文信息，提出了一种具有新型卷积融合层的时空双流网络架构。将光流特征与 RGB 图像信息相结合，双流卷积网络的两个通道通过卷积层获得的特征图进行相互叠加；通过三维（3D）卷积网络将时间信息与空间信息组合，进而从视频中提取潜在信息。使用 UCF101 和 HMDB51 基准数据集对所提出的方法进行了测试，实验结果表明，所提出的双流三维卷积网络模型可以显著提高人体行为识别率。

关键词 人体行为识别；双流卷积网络；三维卷积网络

doi：10.3969/j.issn.1672-9528.2025.05.032

0 引言

人体行为识别旨在研究和理解视频中的人体活动，并自动识别视频或图像序列中的行为类型。由于卷积神经网络 [1]（CNN）在图像分类方面取得了较大进步，许多基于 CNN 的方法已经被提出并应用于视频中的人体行为分类。与图像分类方法相比，基于视频的人体行为识别的一个关键因素是视频中的时间信息。传统的人体行为识别方法主要提取手工制作的局部特征，例如，改进的密集轨迹（iDT）[2]、运动边界直方图（MBH）[3] 等方法已被证明是一种有效的特征表示方法。在人体行为识别领域，卷积网络提取的深度特征性能优于传统手工特征 [4]，融合传统手工和深度特征 [5] 优于应用单一特征。

光流特征是视频分析中运动信息的有效表示，主要用于训练 CNN，包括双流卷积网络和 3D 卷积网络 [6]。然而，提取密集光流仍然是一项耗时的操作，为解决这一问题，Sun 等人 [7] 提出了一种有效地提取光流的方法，即光流引导特征（OFF），可以快速、稳健地提取光流。双流卷积网络将时间和空间网络分开训练并将两个网络提取的二维特征匹配图进行融合作为最终的结果进行输出，在二维特征图上的卷积运算只能提取空间信息，而视频连续帧之间包含的重要运动信息会丢失。为解决这个问题并实现更好的性能，基于 3D CNN 的方法已经变得流行，被应用于大量的图像或视频分类任务。

本文通过在双流卷积网络架构中设计一种融合架构将该网络的空间信息和时间信息结合起来，并且将融合后的特征匹配图采用三维卷积和三维池化进行处理来改进原始双流卷积网络架构。

1 改进的网络架构

视频中包含空间和时间信息。对于空间部分，由单个视频帧表示，承载着视频中描绘的场景和对象信息的外观。时间部分则通过视频帧的运动来传达对象的运动状态。双流卷积网络结构主要用于解决二维卷积网络无法有效提取时间特征的问题。在本文中，利用双流卷积网络的优点，融合不同时刻的 RGB 图像和光流图像的特征图，然后通过三维卷积网络进行处理，以获得更丰富的上下文信息。网络架构如图 1 所示。

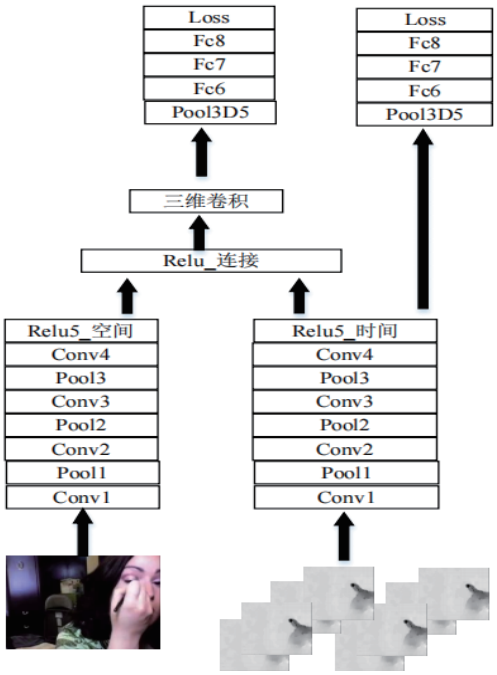


图 1 改进的网络架构图

1. 内江师范学院 四川内江 641199
2. 四川省机场集团有限公司成都天府国际机场分公司 四川成都 641419

[基金项目] 内江师范学院校级科研项目（2024QNZ05）

RGB 图像和 10 个预处理的光流图像分别输入到空间网络和时间网络中。空间网络和时间网络的结果通过卷积操作进行融合，随后通过一个 3D 卷积网络来学习和识别人体行为。该方法的有效性源于以下事实：

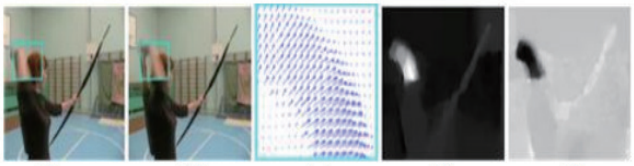
（1）光流图像是运动信息最重要的表示。

（2）不同时间的光流信息和 RGB 图像信息叠加，以获得更全面的运动信息。

（3）使用 3D 卷积网络的目的是在图像中提取更丰富的语义信息，进一步提取全局上下文信息，从而提高人体行为的识别率。

1.1 提取光流图像

稠密光流的计算采用一种逐点匹配图像的配准方法，通过计算图像上所有特征点的偏移量，形成一个密集的光流场，如图 2 所示。



(a) 和 (b) 是相邻两帧图像　(c) 表示 (a) 和 (b) 中方框的光流示意图
(d) 光流的水平方向示意图　(e) 光流的垂直方向示意图

图 2 光流图像示意图

稠密光流可以看作在连续的 t 和 t+1 帧之间的一组位移向量场 d_t。本文用 $d_t(u,v)$ 表示 t 坐标位置 (u,v) 处的位移向量，向量场的水平和垂直部分分别用 d_t^x 和 d_t^y 表示。为代表一系列帧之间的运动，堆叠 L 个连续帧的光流图像形成一个输入通道长度为 $2L$，对于任意帧的一个卷积网络输入块为 I_t=$R^{w\times h\times 2L}$，w 代表视频宽度，h 为视频的高度，公式构建为：

$$I_t(u,v,2k-1)=d^x_{t+k-1}(u,v) \quad (1)$$

$$I_t(u,v,2k)=d^x_{t+x-1}(u,v) \quad (2)$$

对于图像中的任意点 (u,v)，通道 $I_t(u,v,c)$，c=[1:2L] 通过一系列 L 帧图像对该点的动作信息进行编码，运用上述方法提取光流信息后，时间网络的输入是连续 10 帧光流图像。

1.2 RGB 图像和光流图像融合

本文提出的双流三维卷积网络架构采用 ResNet101[8] 深度残差网络来训练空间网络和时间网络，该网络在图像特征提取方面表现较好，并能加速超深层神经网络的训练。

在图像输入阶段，空间网络的输入是由一张张 RGB 图像组成，而时间网络同时输入 10 个不同时刻的光流图像，因此每个行为的 RGB 图像对应于 10 个不同时刻的光流图像。两个网络的输入图像分别进入不同的特征提取层，一张 RGB 图像获得一个 256×7×7 的特征图，其中 256 代表特征图的数目。时间网络输入的是 10 个 x 通道和 10 个 y 通道的光流堆叠，并经过卷积层输出特征匹配图，因此最终时间网络输出的特征图是 10×256×7×7，后取一个空间网络的特征图放在时间网络特征图后面。由于时间网络的输入是通过堆叠 10 个连续的光流图像，而空间网络处理的是单帧 RGB 图像，因此需要复制 10 次空间网络的特征图。将复制的 RGB 特征图与光流特征图叠加，可以得到一个新的特征图，该特征图的结构是一个光流图像特征图叠加一个 RGB 图像特征图。整个视频叠加完成后，输入三维卷积网络中以进一步提取视频的有效信息，流程图如图 3 所示。

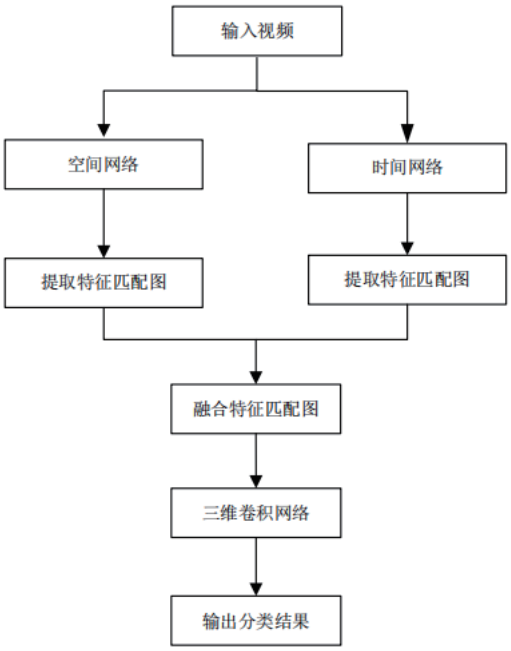


图 3 基于改进双流三维卷积网络的流程图

1.3 三维卷积网络

为从视频中有效捕捉运动信息，Ji 等人 [9] 提出了一种三维卷积来替换原始卷积网络中的二维卷积。三维卷积可以同时捕捉时间和空间信息，使得提取的特征对于动作分类更具辨别力。

为充分利用从时间和空间网络中获取的整个视频信息，本文使用一个三维卷积网络来融合时间网络和空间网络的结果，以捕捉时间和空间维度上具有影响力的有效特征。

三维卷积是通过使用三维卷积核卷积由连续多帧图像组成的空间时间立方体来获得。因此，在连续图像中捕捉运动信息时，第 i 层第 j 个通道的特征图上坐标处的像素值可以通过公式获得：

$$v_{ij}^{xyz}=f\left(b_{ij}+\sum_m\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1}w_{ijm}^{pqr}v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \quad (3)$$

式中：$f(x)$ 表示非线性激活函数，由于修正线性单元具有计算简单、避免梯度消失和梯度爆炸的优点，激活函数成为一种流行的选择；b_{ij} 表示特征图的偏置；m 表示与当前特征图

相连的第 i-1 层的索引；P_i、Q_i 和 R_i 分别表示三维卷积核的高度、宽度和时间维度；w_{ijm}^{pqr} 表示第 m 个特征图与前一层在卷积核图中坐标 (p, q, r) 的权重连接。在池化层中，特征图的像素值是通过池化上一层局部邻域获得的。

卷积网络训练的主要任务是通过对大规模样本的学习，获得卷积层中的卷积核参数 w_{ijm}^{pqr} 和偏置 b_{ij}。在本文中选择了 3×3×3 卷积尺寸作为一个三维卷积网络的卷积核。

2 实验结果和评估

2.1 数据集

本文通过使用 UCF101[10] 和 HMDB51[11] 数据集测试和评估提出的方法，两个数据集是目前评估人体行为识别算法性能的首选数据集，在实验过程中展示训练和测试数据的平均精度。将数据集分为三部分：训练数据、测试数据和计算平均分类准确度的数据。UCF101 总共有 13 320 个视频片段，包括 101 个动作类别。HMDB51 的视频数据来源于电影和公共视频数据库，包括 51 个动作类别，总共包含 6 849 个视频。

2.2 网络训练

本文的网络架构是使用 PyTorch 框架构建的。在空间网络中，首先使用 ResNet101 在 ImageNet 上进行预训练，从视频帧中随机采样 224 px×224 px 大小的图像，再减去每个像素的平均值。使用步长为 2 的下采样卷积层，网络使用全局平均池化，本网络架构的激活函数使用 ReLU，并且在每个卷积层之后添加了局部响应归一化层。使用小批量 256 的随机梯度下降算法，初始学习率为 0.1，当误差稳定时将误差除以 10，并将模型训练最多 6×10^5 次迭代。使用权重衰减为 0.000 1 和动量为 0.9 的参数，并且不使用 dropout 进行处理。

在时间网络中，时间网络的输入数据由 10 个 x 通道和 10 个 y 通道组成。通道图像是由光流图像堆叠而成，输入形状为 (20, 224, 224)，可以看作是一个 20 个通道的图像。为在模型上使用 ImageNet 预训练权重，须将第一个卷积层的预训练权重从 (64, 3, 7, 7) 修改为 (64, 20, 7, 7)。为增加训练样本，未将光流图像裁剪为 224 px×224 px，而是随机地将输入视频帧的宽度和高度上下浮动 25%，在距离图像边界最大 25% 的距离（相对于宽度和高度）处进行裁剪为 224 px×224 px 的大小，并对网络架构进行训练。

3 实验结果和分析

将提出的改进双流卷积网络与其他前沿的传统机器学习方法和经典的双流卷积网络方法进行比较后，本文提出的基于双流网络融合方式的改进算法在 UCF101 和 HMDB51 上的平均识别率分别为 89.1% 和 62.1%，优于目前人工提取特征的方法和经典的双流卷积网络方法。因此将双流卷积网络的空间网络与时间网络的特征进行交互，可以提取更加有效的特征，充分表示整个视频的行为信息，证明了本文提出的网络架构的有效性。如表 1 所示。

表 1 各方法的平均精度

各文献提出的方法	数据集 /%	
	UCF101	HMDB51
iDT	85.9	57.2
iDT with high-dimensional encodings	87.9	61.1
Two-stream with average fusion	86.9	57.9
Two-stream with SVM fusion	88.1	59.4
本文方法	89.1	62.1

4 总结

本文设计了一种新型时空双流三维卷积网络架构，通过在两个网络之间整合一种新颖的卷积融合层，并采用三维卷积进行池化操作，为相邻的输入帧生成了多个信息层，该网络模型能够在图像中提取更丰富的语义信息，并进一步提高全局上下文信息的利用率，从而提高最终的识别率。该模型的性能通过使用 UCF101 和 HMDB51 数据集进行评估，实验结果显示所提出的网络模型优于手工提取特征的方法和经典的双流卷积网络方法，并证明了该网络模型在人体行为识别方面的有效性。未来工作的重点将是考虑如何减少训练时间和计算复杂度，训练更加轻量的网络模型。

参考文献：

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM,2017,60(6):84-90.

[2] WANG H, SCHMID C. Action recognition with improved trajectories[C/OL]// 2013 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2013[2024-05-22]. https://ieeexplore.ieee.org/document/6751553.DOI: 10.1109/ICCV.2013.441.

[3] DALAL N, TRIGGS B, SCHMID C. Human detec-tion using oriented histograms of flow and appearance[C]//Computer Vision–ECCV 2006. Piscataway: IEEE, 2006: 428–441.

[4] WANG L M, QIAO Y, TANG X O. Action recognition with trajectory-pooled deep-convolutional descriptors[EB/OL]. (2015-05-19)[2024-03-13].https://doi.org/10.48550/arXiv.1505.04868.

[5] ZHANG B W, WANG L M, WANG Z, et al.Real- time action recognition with enhanced motion vector CNNs[C/OL]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2016[2024-06-16]. https://ieeexplore.ieee.org/document/7780666.DOI:10.1109/CVPR.2016.297.

[6] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[EB/OL]. (2018-02-12)[2024-09-13].https://doi.org/10.48550/arXiv.1705.07750.

基于非线性干扰观测器的四旋翼无人机容错控制

吕腾飞[1] 苏 彬[2] 钱 宇[3]
LÜ Tengfei SU Bin QIAN Yu

摘 要 为更好地解决四旋翼无人机执行器发生故障时，实现高精度、鲁棒性强且安全飞行的问题，文章结合干扰观测器设计了一种滑模主动容错控制方案。首先，通过设计四旋翼无人机双闭环控制结构，将系统解耦为位置子系统和姿态子系统，结合双曲正切跟踪微分器构建非线性干扰观测器，实时估计并补偿多源外界干扰与执行器故障的集总扰动；其次，基于双幂次趋近律设计滑模容错控制器，以抑制系统抖振并提升收敛速度。仿真实验表明，所提控制策略能表现出更强的鲁棒性和容错性能，验证了其在复杂故障环境下的有效性和优越性，为四旋翼无人机的飞行安全提供了一套可行的控制方案。

关键词 四旋翼无人机；执行器故障；非线性干扰观测器；滑模控制；容错控制

doi：10.3969/j.issn.1672-9528.2025.05.033

0 引言

随着“低空经济”时代的来临，四旋翼无人机展现出广阔的应用前景[1]。然而，其执行器（如电机、螺旋桨）在复杂工况下易受机械磨损、电压扰动或外界冲击影响，导致效率下降或突发性故障，严重威胁飞行稳定性与任务可靠性[2]。如何在执行器故障与多源外界干扰并存场景下实现鲁棒飞行控制，成为四旋翼无人机控制领域的重大挑战。

现有研究大多围绕动力学建模、干扰抑制和容错控制展开。传统方法如 PID 控制依赖线性化模型设计，虽易于工程实现，但对非线性动力学特性与强耦合干扰的适应性不足[3]。近年来，滑模控制（SMC）凭借其强鲁棒性被广泛用于抗干扰设计，但其固有的高频抖振问题易加剧执行器磨损，限制其实际应用[4]。自适应控制可通过在线调节参数补偿模型不确定性，但面临收敛速度与稳态精度之间的权衡难题[5]。在故障容错方面，文献 [6] 提出基于观测器的故障诊断策略，但其依赖精确故障模型，对复合故障的泛化能力有限。

针对多源干扰与执行器故障的协同抑制难题，文献 [7] 采用干扰观测器（DOB）通过估计集总扰动提升系统鲁棒性，

1. 中国民用航空飞行学院计算机学院 四川德阳 618307
2. 中国民用航空飞行学院科研处 四川德阳 618307
3. 中国民用航空飞行学院飞行技术学院 四川德阳 618307
[基金项目] 国家自然科学基金民航联合基金重点项目（U2133209）

[7] SUN S Y, KUANG Z H, SHENG L, et al. Optical flow guided feature: a fast and robust motion representation for video action recognition[C/OL]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE,2018[2024-10-13].https://ieeexplore.ieee.org/document/8578249.DOI: 10.1109/CVPR.2018.00151.
[8] HE K M, ZHANG X Y, REN S Q,et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2016: 770-778.
[9] JI S W, XU W, YANG M, et al.3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013,35(1):221-231.
[10] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[EB/OL].(2012-12-03)[2024-06-23].https://doi.org/10.48550/arXiv.1212.0402.
[11] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]//2011 International Conference on Computer Vision.Piscataway: IEEE, 2011: 2556-2563.

【作者简介】
罗伟 (1994—)，男，四川成都人，本科，研究实习员，研究方向：机器学习、图形图像处理。
陈俊 (1994—)，女，四川成都人，硕士，助理工程师，研究方向：机器学习、人体行为识别。
（收稿日期：2024-12-31 修回日期：2025-04-29）