计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(8):118−125 [doi: 10.15888/j.cnki.csa.007534]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多流卷积神经网络的行为识别①

周　波, 李俊峰

(浙江理工大学 机械与自动控制学院, 杭州 310018)
通讯作者: 李俊峰, E-mail: ljf2003@zstu.edu.cn

摘　要: 人体行为识别与人体姿态有很强的相关性, 由于许多公开的行为识别的数据集并未提供相关姿态数据, 因此很少有将姿态数据进行训练并与其它模态进行融合的识别方法. 针对当今主流基于深度学习的人体行为识别方法采用 RGB 与光流融合的现状, 提出一种融合人体姿态特征的多流卷积神经网络人体行为识别算法. 首先, 用姿态估计算法从包含人的静态图片生成人体关键点数据, 并对关键点连接构建姿态; 其次, 分别将 RGB、光流、姿态数据对多流卷积神经网络进行训练, 并进行分数融合; 最后, 在 UCF101 与 HMDB51 数据集进行了大量的消融, 识别精度等方面的实验研究. 实验结果表明, 融合了姿态图像的多流卷积神经网络在 UCF101 与 HMDB51 数据集的实验精度分别提高了 2.3% 和 3.1%. 实验结果验证了提出算法的有效性.

关键词: 神经网络; 行为识别; 深度学习; 姿态估计; 机器视觉

引用格式: 周波,李俊峰.基于多流卷积神经网络的行为识别.计算机系统应用,2021,30(8):118–125. http://www.c-s-a.org.cn/1003-3254/7534.html

Behavior Recognition Based on Multi-Stream Convolutional Neural Network

ZHOU Bo, LI Jun-Feng

(School of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Human behavior recognition has a strong correlation with human body poses, but many open datasets for behavior recognition do not provide relevant data of poses. As a result, few recognition methods train pose data and fuse with other modalities. Current mainstream behavior recognition methods based on deep learning fuse RGB images with optical flow. This study proposes a behavior recognition algorithm based on a multi-stream convolutional neural network, which integrates human body poses. Firstly, the pose estimation algorithm is used to generate the data of key points on the human body from the static pictures containing people, and the poses are constructed by connecting the key points. Secondly, RGB, optical flow, and pose data are respectively trained on the multi-stream convolutional neural network, and the scores are fused. Finally, substantial experimental research is conducted on ablation and recognition accuracy in UCF101 and HMDB51 datasets. The experimental results reveal that the experimental precision of the multi-stream convolutional neural network integrated with pose images increases by 2.3% and 3.1% in the UCF101 and HMDB51 datasets, respectively, proving the effectiveness of the proposed algorithm.

Key words: neural network; behavior recognition; deep learning; pose estimation; computer vision

行为识别任务旨在从输入的包含行为的视频段中辨别出对应的行为类别. 其应用领域十分广泛, 涵盖了人机交互、智能监控、体育竞技、虚拟现实等领域. 近年来与之相关的研究也受到更多的关注. 受光照条件、相机角度、人体样貌、背景变化等因素影响, 行为识别已成为一项非常具有挑战性的任务.

① 基金项目: 国家自然科学基金 (61374022)
Foundation item: National Natural Science Foundation of China (61374022)
收稿时间: 2020-01-08; 修改时间: 2020-02-08; 采用时间: 2020-02-24; csa 在线出版时间: 2021-07-31

早期经典的行为识别方法主要是基于手工设计的特征. 如 Laptev 等[1] 提出的基于时空兴趣点 (Space-Time Interest Points, STIP) 的方法能在具有较复杂背景的图像的行为识别中取得较好效果. Knopp 等[2] 提出的对 2D-SURF (Speeded Up Robust Features) 特征进行扩展, 形成 3D-SURF 特征, 以期得到更多的 Harr-wavelet 特征. Kläser 等[3] 将局部梯度方向直方图 (Histogram of Oriented Gradient, HOG) 特征扩展到三维形成 HOG-3D, 以在多尺度下对时空块进行快速密度采样. Scovanner 等[4] 将 SIFT 特征 (Scale-Invariant Feature Transform) 扩展到 3D-SIFT. 更多特征描述符还有光流梯度方向直方图 (Histogram of Optical Flow, HOF)[5], 方向梯度直方图 (Histogram of Oriented Gradient, HOG) 及运动边界直方图 (Motion of Boundary History, MBH)[6] 等. 在众多基于手工设计特征的行为识别方法中, IDT (Improved Dense Trajectory)[7] 是目前实验效果最好的方法. 它通过沿着密集轨迹汇集丰富的描述符并补偿相机运动来明确地考虑运动特征, 然后通过编码诸如 BoW, Fisher 矢量的方法, 将描述符聚合到视频级表示中. iDT 在行为识别领域的突出表现使得其与现今基于深度学习的方法仍有结合, 譬如 Wang 等[8] 提出的 TDD (Trajectory-pooled Deep-convolutional Descriptor), 首先, 用深度学习的方法来做特征映射; 其次, 做特征正则化; 最后, 进行轨迹池化操作得到 TDDs.

现今基于深度学习的行为识别方法按输入图像数据的类型可分为: RGB、光流、骨架、帧间差、RGB-D、人体关键点、语音等类型, 以及上述多模态融合的方法. Karpathy 等[9] 提出基于单帧 RGB 图像的深度学习人体行为识别模型. Simonyan 等[10] 提出结合 RGB 图像与光流图像的双流卷积神经网络的方法. 模态融合方法以 RGB 加光流组合较多. 近来在 RGB 与光流组成的双流网络上取得较大突破的有 Wang 等[11] 提出的 TSN (Temporal Segment Networks) 与 Carreira 等[12] 提出的 I3D (Inflated 3D) 网络等. TSN 在双流卷积神经网络的基础上增加了对输入图像进行分段采样的方法, 以期建立能够表征视频级的行为识别模型; I3D 网络由 Google 提出的 Inception 网络改进而来, Inception 网络的卷积层和池化层从 2D 增加一维时间维度变为 3D 的卷积层与池化层, 以期 I3D 网络在保证较深的网络深度与复杂性的同时能够提取视频的时域信息. Sevilla-Luara 等[13] 的研究也证实了将光流模态与 RBG 模态融合可以提升算法在行为识别数据集上的实验精度. 而为了更好地获取视频中的运动信息, Du 等[14] 将 2D-CNN 扩展到 3D-CNN. 为了提升行为识别算法在各数据集上的实验精度, He 等[15] 提出利用 RGB, 光流, 语音信息的多模态行为识别算法.

基于姿态的行为识别与基于 RGB 图像的行为识别相比, 前者的优点是抗干扰能力强, 不易受背景和人的外貌等因素的影响. 因此为了更好的利用姿态在行为识别领域的优势, 受姿态检测算法的启发, 本文提出的算法将姿态检测算法应用在行为识别中, 利用其为行为识别数据集生成关键点特征, 并进行连接, 生成仅包含人体姿态的图像, 并与 RGB 和光流组成 3 种模态融合的模型. 最后在两个主流公开的行为识别数据集上进行了大量实验, 实验结果证明了本文算法的有效性.

1 姿态估计算法

人体姿态数据的获取方式主要分为两种: 其一是通过深度摄像机直接获得, 例如 Kinect, 其二为通过对 RGB 图像中人的关节点进行估计再连接获得.

其中, 第 2 种方法 (姿态估计算法) 又可以分为两类, 一类是自上而下的算法, 另一类是自下而上的算法. 姿态估计算法通过对 RGB 图像中人的关节点进行估计的原理如图 1 所示.

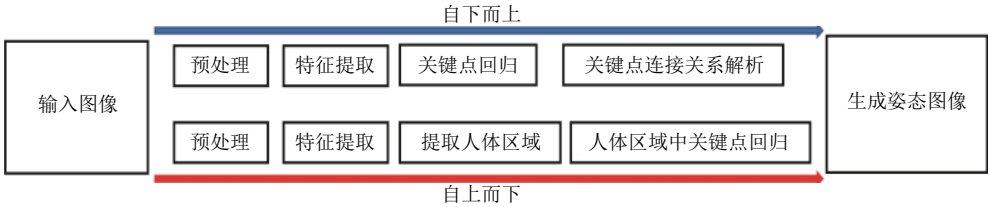


图 1 姿态估计算法原理

本文选取第 2 种方法中的 Openpose[16] 算法对行为识别标准数据集中视频段抽取的每一帧做姿态估计, 获得每一帧中关于人体关键点的热图, 根据热图中的关键点生成每一帧的姿态图. Openpose 在 COCO vali-

dation set 2014, COCO test-dev 2015 以及 COCO 2016 keypoint challenge 的 mAP 分别为 58.4%, 61.8%, 60.5%.

Openpose 属于自下而上的人体姿态估计算法, 与自上向下的算法不同的是, 其先检测人体关键点的位置, 再通过关键点生成骨架. 自上而下的姿态检测算法则是先检测人的区域, 再回归关键点, 该算法的计算效果会直接受影响于前一步骤中人体区域的检测效果, 且计算损耗的时长随着图像中人的个数的增加而增加. 自下而上的算法不受上述两种情况的限制.

Openpose 网络可以大致分为 3 个部分. 第 1 部分为 VGG-19 特征提取网络, 网络的结构如图 2 所示. 输入 $W\times H$ 大小的图像经过预训练的 VGG-19 前 10 层网络得到特征图 F. W 为图像的宽, H 为图像的高.

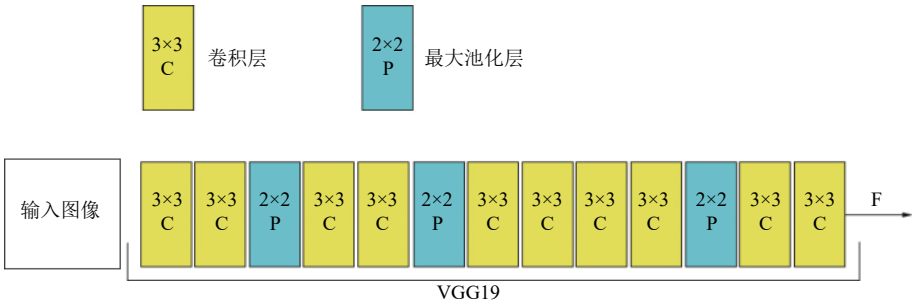


图 2 VGG 特征提取器

输入图像经过了三层池化层, 特征图 F 的大小变为原来的 1/64. 卷积层输出的特征图大小 $W_f\times H_f$ 与原始图像大小 $W\times H$ 及卷积层参数之间的计算公式如下:

$$W_f=\left\lfloor\frac{W-K+2P}{S}+1\right\rfloor \tag{1}$$

$$H_f=\left\lfloor\frac{H-K+2P}{S}+1\right\rfloor \tag{2}$$

其中, W_f 为卷积层输出的特征图的宽, H_f 为卷积层输出的特征图的高, K 卷积核的大小, P 为 *padding* 的值, S 为 *stride* 的值, 池化计算公式与卷积公式基本一致, 池化向上取整, 卷积向下取整.

第 2 部分与第 3 部分的网络结构图如图 2 所示. 特征图 F 经过图 3 左半部分所示的重复的 4 个阶段, 此阶段为人体关键点热图的特征提取部分, 输出为身体各躯干部分亲和力的特征图集 (二维矢量场) G.

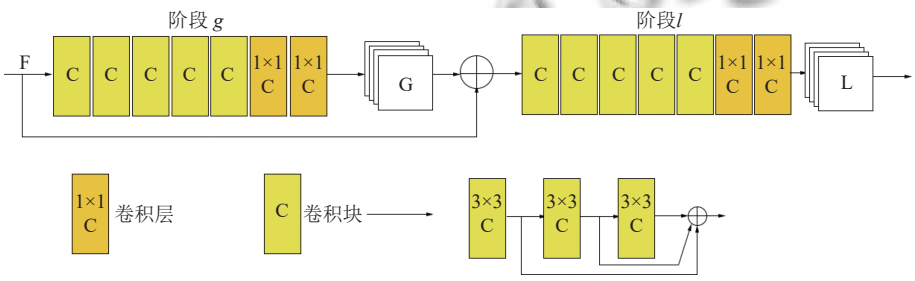


图 3 多阶段 CNN 架构

特征图集 G 与第 1 部分 VGG-19 网络输出的特征图 F 在通道维度叠加后作为第 3 部分的最后 2 个阶段的输入, 该阶段用来生成人体各部分关节点信息的二维置信图集 L.

从二维置信图集 L 解析得到人体关键点, 由二维矢量场 G 可以推测出每一个关键点与其他关键点之间连接关系. 最后的解析步骤执行一组二分匹配以关联身体部位候选关键点, 生成所需的姿态图像. 受数据集的图像质量的影响, 所用姿态估计算法无法在所有待训练与测试的图像中得到令人满意的对应数据, 所以本文实验中对图像中检测到的检测到的人的数量以及每个人体关键点的数量进行评估, 筛选, 留下适合的数

据作为训练与测试的数据. 而未能参与姿态模型测试的图像, 其在多流网络融合中, 姿态网络的得分置 0. 在本文实验中, 融合算法采用平均权重法, 即每个网络的最后得分占最终结果的 1/3. 图 4 显示了 Openpose 算法在人体行为识别数据集 UCF101 与 HMDB51 中的部分图像的姿态估计效果.

图 4 Openpose 处理数据集图像效果图

2 融合姿态数据的多流卷积神经网络人体行为识别算法

本文提出的多流神经网络的总体结构如图 5 所示. 多流神经网络包含 3 个部分, 分别为 RGB 网络, 光流网络与姿态网络.

3 个网络均由 Softmax 层输出预测类别的概率, 最后由 SVM 算法进行最后的得分融合, 得出最终多流神经网络的预测的行为类别.

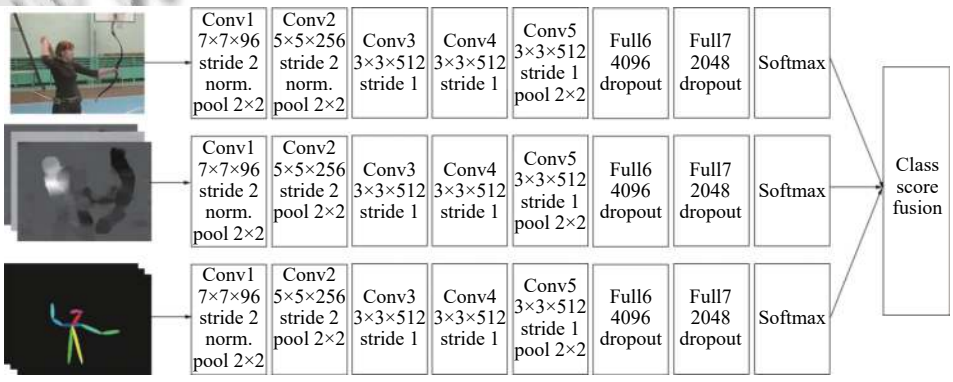


图 5 多流 CNN 体系结构

Softmax 层: 假设卷积神经网络全连接层的输出为 $\gamma_1, \gamma_2, \cdots, \gamma_n$, 这些输出值经由 Softmax 层计算之后会得出一个概率值, 概率值的大小为:

$$softmax(\gamma_i) = \frac{e^{\gamma_i}}{\sum_{j=1}^{n} e^{\gamma_j}} \tag{3}$$

2.1 RGB 网络

视频的单帧 RGB 图像包含视频中所展现的场景信息以及对象信息, 因此从视频中提取的 RGB 图像对行为识别任务具有非常重要的作用. 采用 RGB 图像另一个优点在于相对于光流和 RGB diff 等其他类型数据表示, 静态的 RGB 图像不需要额外处理, 效率较高. 在本文算法中, 利用 RGB 网络来提取视频中图像信息的表观特征. RGB 卷积神经网络由五层卷积层, 两层全连接层组成, 卷积层用来提取特征, 全连接层用来分类, 最后一层为 Softmax 层, 用来对要输出的概率进行归一化. 本文构建的 RGB 流卷积网络采用的是牛津大学视觉几何组 (Visual Geometry Group, VGG) 开发的 VGG-M2048 模型, 网络结构如图 6 所示.

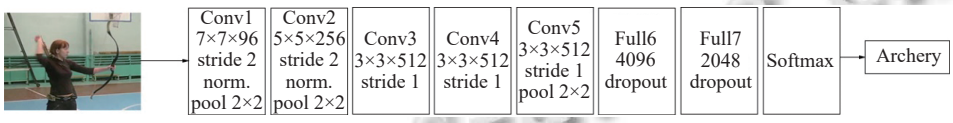


图 6 RGB 网络

2.2 光流网络

光流卷积神经网络结构如图 7 所示, 与 RGB 网络类似, 采用的是 VGG-M-2048 模型. 与 RGB 网的输入不同的是, 光流卷积神经网络输入是按时间发生的前后顺序连续堆叠的光流图像. 光流是分析相邻两帧图像之间像素变化的重要图像亮度模式, 光流图能够被解释为图像中的运动物体的表观运动, 它提供的是图像中像素点的“瞬时速度”, 因此能够更加直观清晰的表征人体运动信息. 光流网络能有效地提取视频中的运动信息, 网络预测类别的得分与 RGB 网络预测类别分数融合, 能显著提高人体行为实验的识别精度.

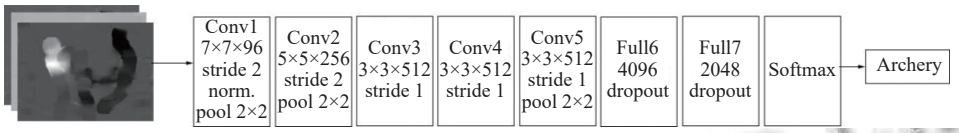


图 7 光流网络

本文实验中的光流帧是采用 OpenCV 中的稠密光流提取法生成的. 光流计算时分别生成了相邻帧的水平方向和垂直方向的光流图像, 将 10 帧水平方向上的光流图与 10 帧垂直方向上的光流图按时间顺序堆叠构成一个光流组, 作为光流卷积神经网络的输入.

2.3 姿态网络

姿态网络可以看作是光流网络的一种扩展. 姿态网络的结构与光流网络的结构相同, 不同的是, 网络的输入从连续堆叠的光流图像变为连续堆叠的姿态图像. 为弥补因相邻帧姿态的相似度高而使姿态网络不能建模长时的姿态信息的不足, 本文对姿态图像进行了分段采样, 再按时间顺序进行堆叠作为姿态网络的输入进行训练. 考虑到不同数量的堆叠姿态图像会对实验结果造成影响, 本文实验部分加入了相关参数的对比实验. 姿态网络结构如图 8 所示.

本文提出的多流卷积神经网络在行为识别任务上不仅考虑了图像的表观信息 (RGB 网络) 与运动信息 (光流网络), 而且融合了行为识别任务中占主导地位的人体的姿态信息 (姿态网络), 最后进行信息融合来互补, 有效地提升行为识别任务的准确率.

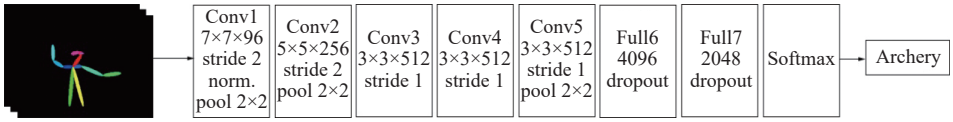


图 8 姿态网络

3 实验结果与分析

3.1 实验数据集

为了便于和目前的主流方法比较, 本文算法在两个比较受欢迎的行为识别数据集上进行了实验, 分别为 UCF101 数据集与 HMDB51 数据集.

UCF101 数据集: 包含 101 个类别的真实场景下的行为识别数据集. 该数据集从 Youtube 网站上取得, 并经过剪辑, 是同为行为识别数据集 UCF50 的扩展. UCF101 数据集中的行为类别共 101 类, 其中每类动作都具有很强的多样性, 这种多样性具体体现在视频中相机的运动, 运动物体的外貌, 形态变化, 背景杂乱等.

HMDB51 数据集: HMDB51 中的视频剪辑主要来源于电影, 剩余部分取自公共数据库和主流的视频网站等. 该数据集包含 6849 个剪辑, 分为 51 个动作类别, 每个动画类别至少包含 101 个剪辑.

3.2 实验条件与结果

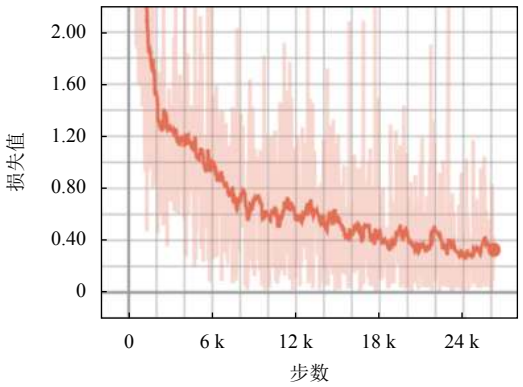
实验条件: 本文算法进行实验过程中的计算机的配置为 Intel Core i5-8500@3.0 GHz, NVIDA GeForce 1080TI GPU, 操作系统为 Ubuntu16.04, 实验框架为 Tensorflow. 训练的超参数上, 学习率设置为 0.001, 动量 0.9, 权重衰减系数为 0.0008, 批量大小设为 16, 使用预训练的双流卷积神经网络, 总的 epoch 大小为 80.

本文使用输入为 224×224 大小图像的神经网络的模型大小为 353 M. 表 1 显示了本文所用模型的以常见的输入图像大小 112×112 为例, 神经网络各层的参数.

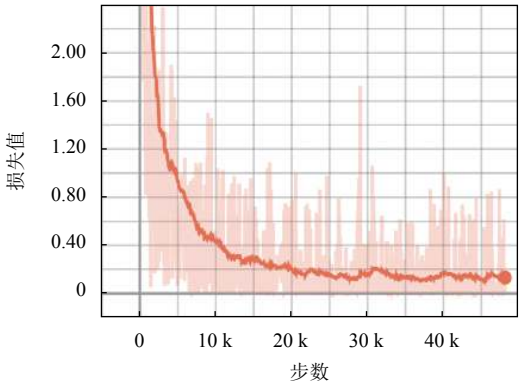
表 1 VGG-2048M 参数

Layer	Output shape	Param
Conv2d_1	(None, 112, 112, 96)	9504
Activation_1	(None, 112, 112, 96)	0
Max_pooling2d_1	(None, 56, 56, 96)	0
Conv2d_2	(None, 28, 28, 256)	614656
Activation_2	(None, 28, 28, 256)	0
Max_pooling2d_2	(None, 14, 14, 256)	0
Conv2d_3	(None, 14, 14, 512)	1180160
Activation_3	(None, 14, 14, 512)	0
Conv2d_4	(None, 14, 14, 512)	2359808
Activation_4	(None, 14, 14, 512)	0
Conv2d_5	(None, 14, 14, 512)	2359808
Activation_5	(None, 14, 14, 512)	0
Max_pooling2d_3	(None, 7, 7, 512)	0
Flatten_1	(None, 25088)	0
Dense_1	(None, 4096)	102764544
Activation_6	(None, 4096)	0
Dropout_1	(None, 4096)	0
Dense_2	(None, 2048)	8390656
Activation_7	(None, 2048)	0
Dropout_2	(None, 2048)	0
Dense_3	(None, 101)	206949

图 9 分别显示了实验过程中姿态网络在 UCF101 数据集 split2 与 HMDB51 数据集训练时的收敛过程. 可见二者分别在 20 k 与 35 k 步数趋稳.



(a) HMDB51



(b) UCF101

图 9 姿态网络在 HMDB51 与 UCF101 的训练收敛特性图

图 10 显示了在两个不同的数据集中, 本文提出的多流网络较双流网络在测试时实验精度上提升最大的行为类别及其提升的效果. 而在本文实验中, 这些类别在姿态估计算法对其进行姿态估计时, 能产生较其它类别更多的训练与测试数据. 而具体在 UCF101 与 HMDB51 中具有姿态训练数据较多且经多流网络融合后实验精度提升最大的 3 个类别分别为 Baby Crawling, Sky Diving, Archery 与 Kick Ball, Swing Baseball, Push Up. 这些类别数据的一个显著的特点是在姿态上与其它数据集具有较强的区分性.

反之, 如果一个类别与另一个或多个类别的姿态相似性较强, 则引入姿态网络会起到反作用. 实验过程中多流较双流网络在数据集 UCF101 与 HMDB51 实验精度上下降精度最多的 3 个类别分别为 Hands Stand Push Ups, Diving, Front Crawl 与 Run, Sit, Climb. 图 11 直观地显示了两种方法在上述类别的精度对比直方图.

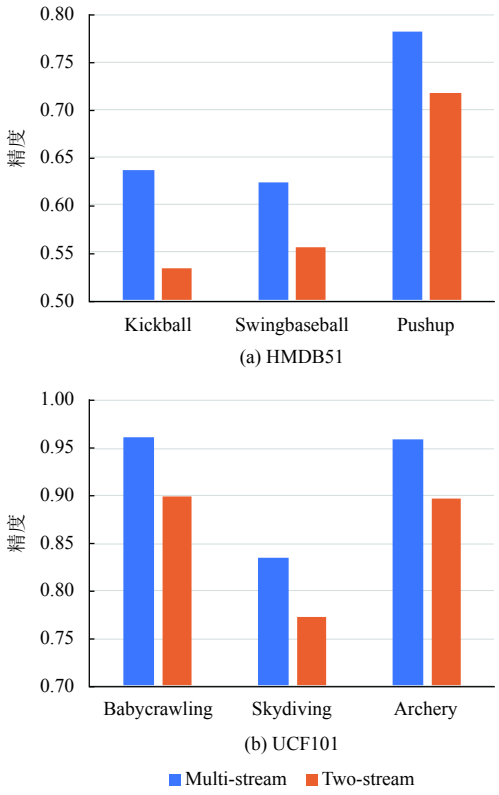


图 10 不同数据集部分类别提升效果直方图

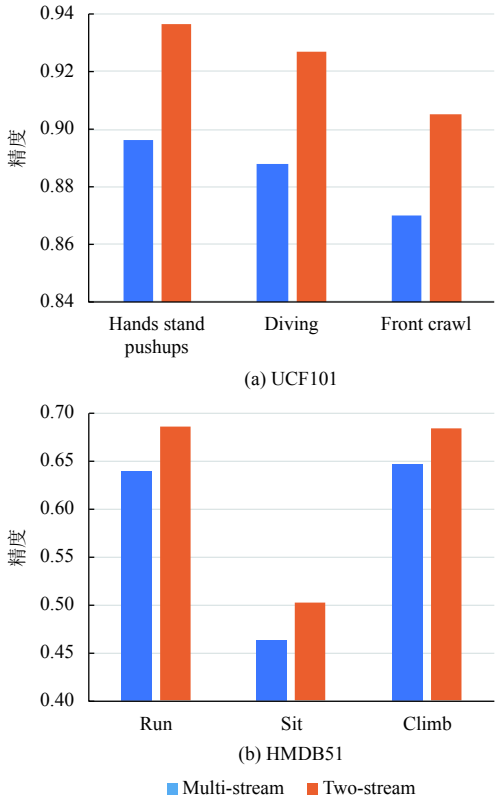


图 11 不同数据集部分类别效果降低直方图

表 2 显示了本文提出算法较原双流卷积算法的消融实验数据对比. 在未融合的双流卷积神经网络的结果上, 为使本文算法的双流网络与原双流卷积神经网络的结果基本一致, 实验中采用预训练的网络, 并且进行了十倍数据增广, 添加了 BN 层等优化操作, 并与双流中的实验结果逐一对比, 最后融合姿态的实验数据. 考虑堆叠的姿态图像的帧数 L 对实验精度影响, 本文选取了从 3 到 10 的不同的堆叠帧数, 并进行了实验, 实验结果如表 2 所示.

表 2 消融实验结果准确度对比 (%)

方法	UCF101	HMDB51
Two stream (RGB)[10]	73.0	40.5
Two stream (Optical Flow)[10]	83.7	54.6
Two stream (Fusion)[10]	88.0	59.4
Ours (RGB)	73.1	40.8
Ours (Optical Flow)	84.8	55.9
Ours (Fusion)	89.3	60.8
Ours (Pose)(L = 3)	47.4	28.1
Ours (Pose)(L = 5)	52.9	31.6
Ours (Pose)(L = 7)	53.1	31.8
Ours (Pose)(L = 10)	53.5	32.1
Ours(Three stream)	**91.3**	**62.5**

由表 3 中的实验结果可知, 融合了姿态图像的多流卷积神经网络在 UCF101 数据集与 HMDB51 数据集上较原方法实验精度分别高出 2.3%, 与 3.1%. 验证了本文提出算法的有效性.

表 3 不同算法在 UCF101 与 HMDB51 数据集上识别准确率对比 (%)

方法	UCF101	HMDB51
IDT[7]	85.9	57.2
MoFAP[17]	88.3	61.7
MIFS [18]	89.1	—
Attention Pooling[19]	—	52.2
C3D (3 nets)[14]	85.2	—
Res3D[20]	85.8	54.9
Two stream[10]	88.0	59.4
Dynamic Image Networks[21]	89.1	—
TDD[8]	90.3	63.2
Ours	**91.3**	**62.5**

4 结论与展望

不同的图像数据类型包含了不同的图像特征, 将不同类型的图像作为深度学习神经网络的训练数据并做融合, 能够提升在单一模态下行为识别的准确度. 具体结论如下:

本文提出了一种结合 RGB、光流、姿态的多流卷积神经网络的行为识别方法, 该方法很好的利用了姿态与行为的强相关性以及在行为表达上的优势, 并且相比于 RGB 图像, 姿态数据更不易受光照, 人的外貌变化等因素的影响. 相较于光流图像, 姿态数据能更好地提供图像种人的形态特征. 实验结果表明, 本文算法在 UCF101 与 HMDB51 数据集上的实验较主流算法有较大提升, 也证实了本文提出的算法的有效性. 接下来的工作是在多模态的基础上实现能够实时运算的方法. 此外, 因光流和姿态图像需要额外的存储空间, 如何将光流和姿态图像数据并入端到端的训练将是实现实时多模态网络的一个关键的待解决的问题.

参考文献

1 Laptev I. On space-time interest points. International Journal of Computer Vision, 2005, 64(2–3): 107–123. [doi: 10.1007/s11263-005-1838-7]

2 Knopp J, Prasad M, Willems G, *et al*. Hough transform and 3D SURF for robust three dimensional classification. Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greek. 2010. 589–602.

3 Kläser A, Marszałek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients. Proceedings of the 19th British Machine Vision Conference. Leeds, UK. 2008. 1−10.

4 Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition. Proceedings of the 15th ACM International Conference on Multimedia. Augsburg, Germany. 2007. 357–360.

5 Laptev I, Marszalek M, Schmid C, *et al*. Learning realistic human actions from movies. Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA. 2008. 1–8.

6 Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance. Proceedings of the 9th European Conference on Computer Vision. Graz, Austria. 2006. 428–441.

7 Wang H, Schmid C. Action recognition with improved trajectories. Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia. 2013. 3551–3558.

8 Wang LM, Qiao Y, Tang XO. Action recognition with trajectory-pooled deep-convolutional descriptors. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 4305–4314.

9 Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 1725–1732.

10 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2014. 568–576.

11 Wang LM, Xiong YJ, Wang Z, *et al*. Temporal segment networks: Towards good practices for deep action recognition. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherland. 2016. 20–36.

12 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6299–6308.

13 Sevilla-Lara L, Liao YY, Güney F, *et al*. On the integration of optical flow and action recognition. Proceedings of the 40th German Conference on Pattern Recognition. Stuttgart, Germany. 2018. 281–297.

14 Du T, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 4489–4497.

15 He DL, Li F, Zhao QJ, *et al*. Exploiting spatial-temporal modelling and multi-modal fusion for human action recognition. https://arxiv.org/abs/1806.10319. (2018-06-27).

16 Cao Z, Simon T, Wei SE, *et al*. Realtime multi-person 2D pose estimation using part affinity fields. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 7291–7299.

17 Wang LM, Qiao Y, Tang XO. MoFAP: A multi-level representation for action recognition. International Journal of Computer Vision, 2016, 119(3): 254–271. [doi: 10.1007/s11263-015-0859-0]

18 Lan ZZ, Lin M, Li XC, *et al*. Beyond Gaussian pyramid: Multi-skip feature stacking for action recognition. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 204–212.

19 Girdhar R, Ramanan D. Attentional pooling for action recognition. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA. 2017. 34–45.

20 Tran D, Ray J, Shou Z, *et al*. ConvNet architecture search for spatiotemporal feature learning. https://arxiv.org/abs/1708.05038. (2017-08-16).

21 Bilen H, Fernando B, Gavves E, *et al*. Action recognition with dynamic image networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2799–2813. [doi: 10.1109/TPAMI.2017.2769085]