基于双流快速区域卷积神经网络改进的人体动作识别算法

郭如意，金杰*，刘高华，刘凯燕，姜诗祺
天津大学电气自动化与信息工程学院，天津 300072

摘要 深度神经网络在静态图像领域已取得突破性进展，并逐步扩展到视频识别领域。人体动作识别是视频识别领域的研究热点和难点，因此，提出了一种基于双流快速区域卷积神经网络(Faster RCNN)改进的人体动作识别算法。首先，用 RGB(Red，Green，Blue)图像和光流数据作为网络的输入，分别训练 Faster RCNN；然后，将训练好后的网络模型进行融合，并引入改进的压缩和激励模块对特征通道进行处理，以突出重要特征；最后，用完全的交并比损失函数作为边框回归损失函数，以优化某些预测框与真实框不能相交等问题。实验结果表明，相比传统的 Faster RCNN，本算法在动作识别数据集 UCF101 上的准确率得到了一定的提高。

关键字 机器视觉；双流快速区域卷积神经网络；人体动作识别；压缩与激励；交并比损失函数
中图分类号 TP391 **文献标识码** A **doi**：10.3788/LOP57.241506

Improved Human Action Recognition Algorithm Based on Two-Stream Faster Region Convolutional Neural Network

Guo Ruyi, Jin Jie*, Liu Gaohua, Liu Kaiyan, Jiang Shiqi
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract In the field of static image, deep neural networks have made breakthroughs and gradually expanded to the field of video recognition. Human action recognition is a research hotspot and difficult in the field of video recognition. Therefore, this paper proposes an improved human action recognition algorithm based on two-stream faster region convolutional neural network (Faster RCNN). First, we use RGB (Red, Green, Blue) images and optical flow data as input of the network to train the Faster RCNN separately; then, the trained network model is fused, and an improved squeeze and excitation block is introduced to process the feature channel to highlight important features; finally, we use the complete intersection-over-union loss function as the bounding box regression loss function to optimize some problems such as the inability to intersect the ground truth box with the predicted box. The experimental results show that the accuracy of the algorithm on the action recognition data set UCF101 is improved compared to the traditional Faster RCNN.

Key words machine vision; two-stream faster region convolutional neural network; human action recognition; squeeze and excitation; intersection-over-union loss function
OCIS codes 150.1135; 100.3008; 040.7290

1 引 言

人体动作识别是视频识别领域的热门课题，根据研究方法可分为传统学习方法和深度学习方法。传统学习方法可分为基于模板匹配[1]、基于时空兴趣点[2]和基于轨迹[3]的方法，深度学习方法可分为基于三维卷积神经网络(C3D)[4]、基于时间卷积网络(TCN)[5]和基于双流卷积神经网络(CNN)[6]的方法。

基于模板匹配的方法用一组模板表示动作，求解需要识别的动作模板和已有模板之间的相似度，当两个模板之间的相似度小于预先设定的阈值时，

收稿日期：2020-04-29；**修回日期**：2020-05-27；**录用日期**：2020-06-17
基金项目：国家自然科学基金(61571320)
*E-mail：jinjie@tju.edu.cn

就能得到判定结果。咎宝锋等[7]提出了一种判别协作表征分类器，考虑了样本对协作表征系数的影响，提升了对相似样本的判别精度，但该方法计算量大，且对方向和位置比较敏感。基于时空兴趣点的方法中时空兴趣点是视频中动作变化最剧烈的位置，刘帆等[8]融合了加速稳健特征（SURF）和方向梯度直方图（HOG），同时使用背景减法，消除了无关背景带来的影响。该方法通过兴趣点提取特征信息，可在一定程度上提高准确率，但过于依赖人体轮廓，计算量较大。基于轨迹的方法用人体骨骼关键点表示动作，如密集轨迹（DT）[9]和改进的 DT（IDT）[10]算法。DT 算法通过多尺度提取图像的特征，然后将特征的光流中值作为轨迹；IDT 算法改进了光流图像，进一步提高了动作识别的准确率，同时也克服了相机角度对识别效果的影响。

C3D 可以捕获空间和时间信息，且网络结构简单，由 5 个卷积层、5 个池化层和 2 个全连接层组成，泛化能力较好，但无法长期获取时间信息；TCN 通过引入膨胀卷积[11]解决了 C3D 无法长期获取时间信息的问题，但其迁移能力较差、感受野不够大；双流 CNN 用 RGB（Red，Green，Blue）图像和光流数据作为空间和时间信息，相比上述算法，网络的整体性能得到了一定的提升。黄友文等[12]将 CNN 与长短时记忆（LSTM）网络相结合，实现了浅层特征与深层特征的连接，进一步提高了动作识别的准确率。Shrivastava 等[13]将在线难例挖掘（OHEM）与快速区域卷积神经网络（Faster RCNN）相结合，提高了双流网络检测难例的准确率，但难例挖掘比较困难，需要迭代模型进行训练。Peng 等[14]使用双流 Faster RCNN 分别提取空间信息和时间信息，但准确率较低。Gkioxari 等[15]使用一个包含人、动作和物体的三元素表示人和物体之间的交互关系，提高了动作识别的准确率。

本文使用双流 Faster RCNN[16]进行动作识别，首先，在双流 Faster RCNN 的骨干网络中，引入压缩与激励模块（SE block）[17]，从特征通道的角度自适应获取特征响应值，突出重要特征，可在减少参数和计算量的同时增强网络的泛化能力。其次，改进了 SE block，用 h-swish[18]激活函数代替 Sigmoid 激活函数，进一步提升了网络的性能。最后，用完全的交并比（IOU）损失函数[19]代替 IOU 函数，作为边框回归损失函数，解决了预测框和真实框不相交时无法优化的问题，同时也能确定两个框之间的相对位置关系，进一步提升了动作识别的准确率。

2 网络结构与算法原理

本算法需对视频进行两次处理，第一次处理是将视频分解成 RGB 帧图像，第二次处理是从视频中获得对应的光流信息。算法的整体网络结构如图 1 所示，可以发现，网络由两部分构成，第一部分用来提取空间信息，输入数据为帧图像；第二部分用来提取时间信息，输入数据为光流数据，然后将空间信息与时间信息进行融合后，获得最终结果。其中，RPN 为区域生成网络，CIOU 为完全的交并比，improved SE-VGG16 为基于压缩和激励（SE）改进的视觉几何组（VGG16）网络，cls_prob 为最终的识别分数。

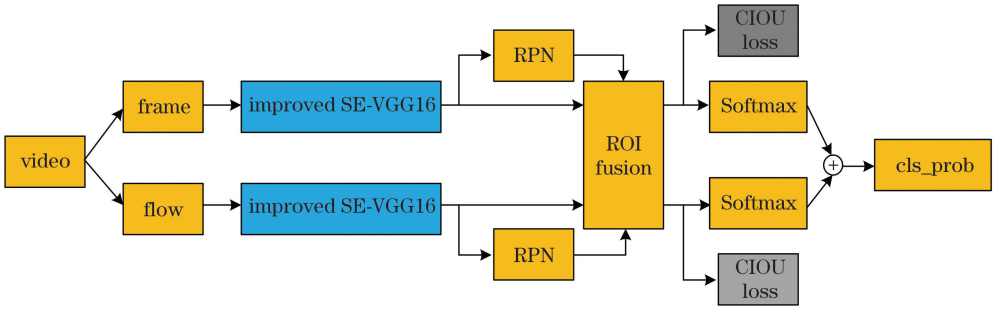

图 1 本网络的整体结构

Fig. 1 Overall structure of our network

2.1 Improved SE-VGG16 的结构

骨干网络作为网络后续的输入，其特征提取的好坏直接影响了后续的检测效果。虽然传统的双流 Faster RCNN 检测效果较好，但并没有重视特征通道，对特征的提取也不充分。为了解决该问题，在传统双流 Faster RCNN 的基础上引入 SE block，网络结构如图 2 所示。SE block 由 1 个全局平均池化（GAP）层、2 个全链接（FC）、1 个修正线性单元

(ReLU)激活函数和 1 个 Sigmoid 激活函数组成。其工作原理可以分为特征压缩和获取自适应权重两部分。假设提取的特征维度为 $w\times h\times c$,压缩后的特征维度为 $1\times1\times c/r$,其中,r 为压缩系数。自适应权重过程将特征维度变为 $1\times1\times c$,并将归一化后的权重赋加到每个特征的通道上,从而突出重要特征。

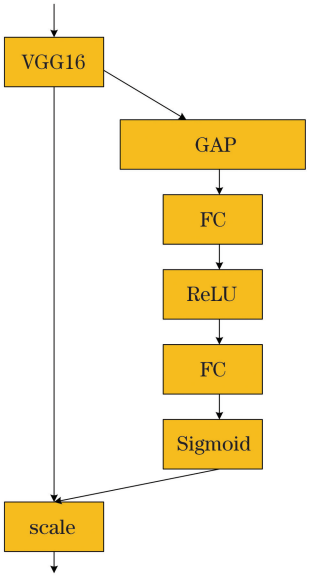

图 2 SE-VGG16 的结构

Fig. 2 Structure of SE-VGG16

由于 Sigmoid 激活函数计算复杂,求解不准确,且收敛速度较慢。因此,用 h-swish 激活函数替换 Sigmoid 激活函数,其网络结构如图 3 所示。h-swish 激活函数可表示为

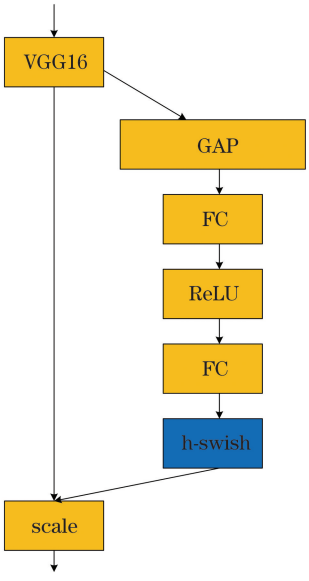

图 3 改进 SE-VGG16 的结构

Fig. 3 Structure of Improved SE-VGG16

$$A_{\text{h-swish}}(x)=x\frac{A_{\text{ReLU6}}(x+3)}{6},\tag{1}$$

式中,x 为输入特征,A_{ReLU6} 为 ReLU 激活函数的变体,可加快收敛速度。

$$A_{\text{ReLU6}}(x)=\min[\max(x,0),6],\tag{2}$$

$$A_{\text{ReLU6}}(x+3)=\min[\max(x+3,0),6]。\tag{3}$$

2.2 边框回归损失函数

传统双流 Faster RCNN 的边框损失函数为 IOU 损失函数,可表示为

$$L_{\text{IOU}}=-\ln\left(\frac{G\cap P}{G\cup P}\right),\tag{4}$$

式中,G 为真实框,P 为预测框。可以发现,当真实框与预测框不相交,即 $G\cap P=0$ 时,损失函数不可导,无法进行优化;当真实框与预测框相交时,损失函数为定值,无法确定两个框的相对位置关系。针对这些问题,用 CIOU 作为双流 Faster RCNN 中的损失函数,可表示为

$$L_{\text{CIOU}}=1-\frac{G\cap P}{G\cup P}+\frac{d^2(O_{\text{p}},O_{\text{G}})}{C^2}+\alpha V,\tag{5}$$

式中,O_{p}、O_{G} 分别为预测框和真实框的中心点,d 为预测框和真实框之间的欧氏距离,C 为预测框和真实框最小外接矩形的对角线距离,V 为衡量预测框和真实框宽高比相似性的参数,α 为权重函数,可表示为

$$V=\frac{4}{\pi^2}\left(\arctan\frac{w_{\text{G}}}{h_{\text{G}}}-\arctan\frac{w_{\text{p}}}{h_{\text{p}}}\right)^2,\tag{6}$$

$$\alpha=\frac{V}{1-\frac{G\cap P}{G\cup P}+V},\tag{7}$$

式中,w_{G}、h_{G} 分别为真实框的宽度、高度,w_{p}、h_{p} 分别为预测框的宽度、高度。由(5)式可知,当预测框和真实框不相交时,可对其梯度进行优化;当预测框和真实框相交时,可根据预测框和真实框的中心点 O、宽度 w 和高度 h 确定两个框的相对位置,从而解决 IOU 损失函数的优化问题。

3 实验与结果分析

3.1 实验设置

实验环境:深度学习框架为 Caffe,计算机操作系统为 Ubuntu 16.04 LTS,GPU 的型号为 GTX 1080 Ti(2 块)。用 UCF101 数据集进行训练测试,UCF101 数据集中图像的分辨率为 320 pixel×240 pixel,共有 101 类动作,每类动作被分成 25 组,每组约包含 6 个视频。选取其中 24 类动作进行实验,分别为 basketball, basketball dunk, biking,

cliff diving, cricket bowling, diving, fencing, floor gymnastics, golf swing, horse riding, ice dancing, long jump, pole vault, rope climbing, salsa spin, skate boarding, skiing, ski jet, soccer juggling, surfing, tennis swing, trampoline jumping, volleyball spiking, walking with dog。在每类动作中,将第 1 组到第 7 组拍摄的视频作为测试集,共 914 个视频,第 8 组到第 25 组拍摄的视频作为训练集,共 2293 个视频。初始化学习率为 0.001,进行 100000 次迭代后,学习率减为 0.0001,140000 次迭代后,停止训练。

3.2 实验结果

不同算法在 UCF101 数据集上的平均精度均值(mAP)如表 1 所示,可以发现,相比 IDT、C3D、轨迹池深度卷积描述符(TDD)[20]、Two-stream 算法,本算法的 mAP 分别提高了 6.4、7.6、2.5、4.8 个百分点。但相比时空残差网络(ST-ResNet)算法[21],本算法的准确率还存在一定差距,原因是 ST-ResNet 算法的骨干网络为残差网络(ResNet),比本算法的骨干网络更深,具有更好的特征提取能力,但 ST-ResNet 算法的参数量较多,因此本算法的性能更优越。

表 1 不同算法在 UCF101 数据集上的 mAP

Table 1 mAP of different algorithms on UCF101 data set

unit: %

Method	mAP
IDT	86.4
C3D	85.2
TDD	90.3
Two-stream	88.0
ST-ResNet	93.4
Ours	**92.8**

表 2 为不同模块对算法准确率的影响,可以发现,相比传统双流 Faster RCNN,引入 SE block,能突出重要特征,使准确率提高了 1.6 个百分点;使用改进的 SE block,可使准确率提高 2.0 个百分点;在此基础上使用 CIOU 损失函数,可解决 IOU 损失函数在预测框与真实框不相交时无法优化的问题,使准确率提高 2.3 个百分点。图 4 为本算法在不同场景下的检测结果。

表 2 不同模块对本算法的影响

Table 2 Influence of different modules on our algorithm

unit: %

Method	VGG16	SE-VGG16	Improved SE-VGG16	IOU loss	CIOU loss	mAP
Two-stream Faster RCNN	√			√		90.5
Two-stream Faster RCNN		√		√		92.1
Two-stream Faster RCNN			√	√		92.5
Two-stream Faster RCNN			√		√	**92.8**

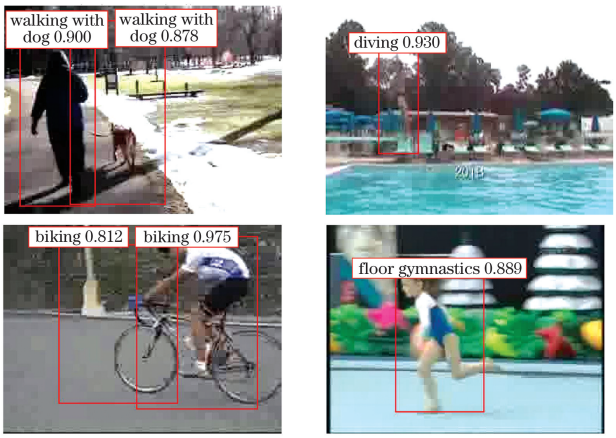

图 4 本算法的检测结果。(a)遛狗;(b)跳水;(c)骑自行车;(d)体操

Fig. 4 Detection results of our algorithm. (a) Walking with dog; (b) diving; (c) biking; (d) floor gymnastics

4 结 论

在传统双流 Faster RCNN 的基础上,引入了 SE block,突出了重要特征;同时,改进了 SE block,用 h-swish 函数代替 Sigmoid 函数,用 CIOU 损失函数代替 IOU 损失函数,解决了预测框与真实框不相交时无法优化以及相交时相对位置定位不确定的问题。实验结果表明,相比原始算法,改进后的算法动作识别准确率更高。

参 考 文 献

[1] Wang H Y, Qi J, Fang T E, et al. Dynamic hand gesture recognition based on track template matching [J]. Microcontrollers & Embedded Systems, 2017, 17(7): 39-43, 46.
王浩宇, 漆晶, 方天恩, 等. 基于轨迹模板匹配的动态手势识别方法[J]. 单片机与嵌入式系统应用,

2017, 17(7): 39-43, 46.

[2] Chen S D, He B Q, Chen S Y, et al. Human action recognition based on spatio-temporal interest point [J]. Journal of Chengdu University of Information Technology, 2018, 33(2): 143-148.
陈胜娣, 何冰倩, 陈思宇, 等. 基于时空兴趣点的人体动作识别[J]. 成都信息工程大学学报, 2018, 33(2): 143-148.

[3] Dong S G, Hu D D, Li R J, et al. Human action recognition based on foreground trajectory and motion difference descriptors[J]. Applied Sciences, 2019, 9(10): 2126.

[4] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 4489-4497.

[5] Lea C, Flynn M D, Vidal R, et al. Temporal convolutional networks for action segmentation and detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1003-1012.

[6] Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal residual networks for video action recognition[EB/OL]. [2020-04-15]. https://arxiv.org/abs/1611.02155.

[7] Zan B F, Kong J, Jiang M. Human action recognition based on discriminative collaborative representation classifier[J]. Laser & Optoelectronics Progress, 2018, 55(1): 011010.
咎宝锋, 孔军, 蒋敏. 基于判别协作表征分类器的人体行为识别[J]. 激光与光电子学进展, 2018, 55(1): 011010.

[8] Liu F, Yu F Q. Human action recognition based on global and local features[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021004.
刘帆, 于凤芹. 基于全局和局部特征的人体行为识别[J]. 激光与光电子学进展, 2020, 57(2): 021004.

[9] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition [J]. International Journal of Computer Vision, 2013, 103(1): 60-79.

[10] Wang H, Schmid C. Action recognition with improved trajectories[C]//2013 IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, NSW, Australia. New York: IEEE, 2013: 3551-3558.

[11] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[12] Huang Y W, Wan C L, Feng H. Multi-feature fusion human behavior recognition algorithm based on convolutional neural network and long short term memory neural network[J]. Laser & Optoelectronics Progress, 2019, 56(7): 071505.
黄友文, 万超伦, 冯恒. 基于卷积神经网络与长短期记忆神经网络的多特征融合人体行为识别算法[J]. 激光与光电子学进展, 2019, 56(7): 071505.

[13] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[EB/OL]. [2020-04-13]. https://arxiv.org/abs/1604.03540.

[14] Peng X J, Schmid C. Multi-region two-stream R-CNN for action detection[M]//Leibe B, Matas J, Sebe N, et al. Computer Vision-ECCV 2016. Lecture Notes in Computer Science. Cham: Springer, 2016, 9908: 744-759.

[15] Gkioxari G, Girshick R, Dollár P, et al. Detecting and recognizing human-object interactions[EB/OL]. [2020-04-13]. https://arxiv.org/abs/1704.07333.

[16] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[17] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[EB/OL]. [2020-04-12]. https://arxiv.org/abs/1709.01507.

[18] Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3 [EB/OL]. [2020-04-14]. https://arxiv.org/pdf/1905.02244.

[19] Zheng Z H, Wang P, Liu W, et al. Distance-IoU loss: faster and better learning for bounding box regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993-13000.

[20] Wang L M, Qiao Y, Tang X O. Action recognition with trajectory-pooled deep-convolutional descriptors [EB/OL]. [2020-04-15]. https://arxiv.org/abs/1505.04868.

[21] Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal residual networks for video action recognition [EB/OL]. [2020-04-11]. https://arxiv.org/abs/1611.02155.