计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190180
57(1):145-158,2020

基于特征融合时序分割网络的行为识别研究

李洪均[1,2,3,4]　丁宇鹏[1]　李超波[1]　张士兵[1,3]
[1](南通大学信息科学技术学院　江苏南通　226019)
[2](计算机软件新技术国家重点实验室(南京大学)　南京　210023)
[3](南通智能信息技术联合研究中心　江苏南通　226019)
[4](通科微电子学院　江苏南通　226019)
(lihongjun@ntu.edu.cn)

Action Recognition of Temporal Segment Network Based on Feature Fusion

Li Hongjun[1,2,3,4], Ding Yupeng[1], Li Chaobo[1], and Zhang Shibing[1,3]
[1](*School of Information Science and Technology, Nantong University, Nantong, Jiangsu* 226019)
[2](*State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing* 210023)
[3](*Nantong Research Institute for Advanced Communication Technologies, Nantong, Jiangsu* 226019)
[4](*Tongke School of Microelectronics, Nantong, Jiangsu* 226019)

Abstract Action recognition is a research hot topic and a challenging task in the field of computer vision nowadays. Action recognition analysis is closely related to its network input data type, network structure and feature fusion. At present, the main input data of action recognition network is RGB images and optical flow images, and the network structure is mainly based on two-stream and three dimension convolution. While the selection of features directly affects the efficiency of recognition and there are still many problems to be solved in multi-layer feature fusion. In view of the limitation of the RGB images and optical flow images which are the input of the popular two-stream convolution network, using sparse features in low rank space can effectively capture the information characteristics of moving objects in video and supplement the network input data. Meanwhile, for the lack of information interaction in the deep network, the high-level semantic information and the low-level detailed information are combined to recognize actions together, which makes temporal segment network performance more advantageous. Extensive experiments in subjective and objective comparison are performed on UCF101 and HMDB51 and the results show that the proposed algorithm is significantly better than several state-of-the-art algorithms, and the average accuracy rate of the proposed algorithm reaches 97.1% and 76.7%. The experimental results show that our method can effectively improve the recognition rate of action recognition.

收稿日期:2019-03-21;**修回日期**:2019-07-22

基金项目:国家自然科学基金项目(61871241);教育部产学研合作协同育人基金项目(201802302115);中国交通教育研究会教育科学研究课题(交教研1802-118);南通市科技计划资助项目(JC2018025,JC2018129);南京大学计算机软件新技术国家重点实验室基金项目(KFKT2019B015);江苏省研究生科研与实践创新计划项目(KYCX19_2056);南通大学-南通智能信息技术联合研究中心基金项目(KFKT2017B04)

This work was supported by the National Natural Science Foundation of China (61871241), the Ministry of Education Cooperation in Production and Education (201802302115), the Educational Science Research Subject of China Transportation Education Research Association (Jiaotong Education Research 1802-118), the Science and Technology Program of Nantong (JC2018025, JC2018129), the Nanjing University State Key Laboratory for Novel Software Technology (KFKT2019B015), the Postgraduate Research and Practice Innovation Program of Jiangsu Province (KYCX19_2056), and the Nantong University-Nantong Joint Research Center for Intelligent Information Technology (KFKT2017B04).

Key words action recognition; sparse features; temporal segment network; two-stream convolution network; feature fusion

摘 要 行为识别是当今计算机视觉领域的一个研究热点,是一项具有挑战性的任务.行为识别分析与其网络输入数据类型、网络结构、特征融合环节具有密切联系.目前,主流的行为识别网络输入数据为RGB图像和光流图像,网络结构主要以双流和3D卷积为主;而特征选择直接影响到识别的效率,多层次的特征融合工作还有很多问题有待解决.针对主流的双流卷积网络输入数据为RGB图像和光流图像的局限,利用低秩空间中稀疏特征能够有效捕捉视频中运动物体信息的特点,对网络输入数据进行补充.同时,针对网络中缺乏信息交互的特点,将深度网络中高层语义信息和低层细节信息结合起来共同识别行为动作,使时序分割网络性能更具优势.在行为识别数据集UCF101和HMDB51上取得了97.1%和76.7%的识别效果,较目前主流算法有了较大的提升.实验结果表明,该方法能够有效地提高行为识别的识别率.

关键词 行为识别;稀疏特征;时序分割网络;双流卷积网络;特征融合

中图法分类号 TP391

人体行为识别是一项具有挑战性的任务,受光照不同、背景复杂、多视角、类内差异大等诸多因素的影响[1-3].人体行为识别算法主要分为2种:1)基于传统机器学习的方法[4-13];2)基于深度学习的方法[14-18].这2种方法各有优劣,基于传统机器学习的行为识别算法关键在于特征的提取,研究过程中往往会花费心力设计满足需求的特征,实现简单,但其表征行为动作的能力也受限于提取的特征;基于深度学习的行为识别算法能够自动学习特征,但需要大量数据支撑,自动提取的特征是否有效与网络结构设计、网络参数选取等息息相关.

行为识别中应用深度学习最直接的方法即使用卷积神经网络(convolutional neural network, CNN)对视频的每一帧进行识别,但这种方法并没有考虑到连续视频帧之间的运动信息.Ji等人[19]首次提出了3D卷积的概念,利用3D卷积核提取空间和时间特征用于行为识别.Simonyan等人[20]提出了双流卷积神经网络用于行为识别,该网络分为空间流卷积网络和时间流卷积网络2个部分.空间流卷积网络以单帧RGB图像为输入,表示视频中某一刻的静态表观信息;时间流卷积网络以连续几帧光流图像堆叠在一起为输入,表示物体的运动信息,最后将2个网络的分类结果融合得到最终准确率,该模型的提出打破了改进版的稠密轨迹提取算法(improved dense trajectories, IDT)[21]在行为识别领域的领先地位.Tran等人[15]提出了一种新的3D卷积神经网络(convolutional 3 dimention, C3D),C3D网络将连续视频帧堆叠起来作为网络输入,利用3D卷积核在堆叠后形成的立方体中进行卷积,较2D卷积核多了时间维度,这样可以从连续帧上获取运动信息,该算法最大的优势是识别速度较双流算法提升了很多.至此,行为识别算法形成了两大主流流派:一种是基于双流卷积神经网络的行为识别算法;另一种是基于3D卷积神经网络的行为识别算法.

目前,主流的行为识别网络输入数据为RGB图像和光流图像.对于空间流卷积网络,输入数据为RGB图像,最开始的空间流网络采用逐帧输入的方式,而目前公开的数据集往往单帧图像就能完成识别任务,这种情况下空间流卷积网络的输入就存在大量冗余信息.为了减少逐帧输入时连续帧之间的冗余,Zhu等人[22]提出了一种关键帧获取的方法,挖掘视频中对于行为识别有决定性的帧和关键区域,以此来提升准确率与效率.虽然这个提取关键帧的方法可以集成到1个网络中训练,但是其与目标检测网络RCNN类似,先提取候选框,再选关键帧,网络结构复杂;Kar等人[23]提出了一种AdaScan特征聚集方法,判断不同帧的重要程度,并据此聚集特征以实现提升准确率与效率的目的,该方法整体模型较前一种方法简单.对于时间流卷积网络,输入数据为光流图像,光流提取耗时耗力,并且光流所包含的运动特征未必是最优特征.不少研究者对光流进行了改进,并且对其在行为识别中起到的作用进行了研究.Zhu等人[24]提出了一种双流卷积网络,在时间流网络之前加入了MotionNet生成光流图像,作为时间流卷积网络的输入,该方法提升了光流质量;Sevilla-Lara等人[25]通过实验证明了光流对于行为识别有效是因为它的表观特征不变性,其本身质量评判指标终点误差(end-point-error, EPE)与行为

识别准确率并无强相关性,从测试的光流算法来看,光流在边界处以及小位移处的精度对于行为识别算法性能的提升有强相关性,并且通过行为识别的损失函数值对光流进行改进,使得识别准确率得以提升.同样,由于光流图像的弊端,也有不少研究者在寻找能够替代光流的特征方面做了一些工作.Zhang等人[26]利用运动向量来替代光流,运动向量原本用于视频压缩,不需要额外的计算就可以直接提取,极大地加快了双流卷积网络的识别速度,但精度有所降低;Choutas 等人[27]提出了一种新型姿态特征,通过提取人体关键关节点的轨迹,并对其进行颜色编码,形成姿态特征图像用于行为识别,其对于 RGB 图像和光流图像所提供的特征具有补充作用,单一使用表现不佳.仅通过改变双流网络的交互方式和提取新的运动特征作为网络输入,并不能同时解决精度与速度的问题,网络结构的改变对于算法性能的提升也有决定性的作用.

近年来,主要的行为识别网络结构大都基于双流网络和 3D 卷积网络发展而来.Wang 等人[28]提出了时序分割网络(temporal segment network, TSN),利用多个双流网络提取不同时序位置上的短时运动信息并进行融合,以解决传统双流只关注表观特征和短时运动信息的问题.Lan 等人[29]继承了 TSN 的优良特性,对于不同时序位置上的短时运动信息进行了加权融合;Zhou 等人[30]提出了时序推理网络,该网络建立在 TSN 基础之上,增加了 3 层全连接网络学习不同长度视频帧的权重,并对不同长度的视频帧进行时序推理,最后进行融合得到结果.Xu 等人[31]结合了 C3D 和 Faster-RCNN(faster-region convolutional neural network)[32]的思想提出了 R-C3D(region-convolutional 3D network),R-C3D 使用 3D 卷积提取视频特征,采用了 Faster-RCNN 形式的思路,即先生成提议,再进行候选区域池化,最后进行分类和边界回归,该网络可以对任意长度的视频进行端到端行为识别,并且速度快、通用性好;Qiu 等人[33]针对行为识别中采用的 3D 卷积进行改造,提出了 P3D 网络(pseudo-3D residual net, P3D ResNet),利用 1×1×3 卷积和 3×1×1 卷积代替 3×3×3 卷积,前者与 2D 卷积类似,提取空间流特征,后者用来获取时间流特征,这种方法大大减少了计算量.不仅双流卷积网络和 3D 卷积神经网络可以提取时间流信息,利用长短时记忆网络(long short-term memory, LSTM)[34]也可以进行时间维度建模,这也是目前行为识别领域比较流行的一个方向.Long 等人[35]提出了一种结合注意力机制的多模态 LSTM 结构,稳定性高;Du 等人[36]引入姿态注意力机制,结合了 LSTM 和 CNN 结构,能够有效提取时空特征.另外,还有研究者在构成深度网络的通用部件方面做了研究,Wang 等人[37]提出了一种新型的非局部网络结构,将非局部操作作为一个高效、简单、通用的组件,能够用来捕捉神经网络中的长距离依赖关系.深度学习算法以双流结构和 3D 卷积为主,其中基于双流结构的算法精度高,速度较慢;而基于 3D 卷积的算法速度快,精度略低,整体高于传统的机器学习算法,在应对复杂背景、类内变化大等问题方面较传统算法有很大优势.

本文针对主流的双流卷积网络输入数据为 RGB 图像和光流图像的局限,利用低秩空间中稀疏特征能够有效捕捉视频中运动物体的信息特点,对网络输入数据进行补充.同时,针对网络中缺乏信息交互的特点,将深度网络中高层语义信息和低层细节信息结合起来共同识别行为动作,使网络性能更具优势.本文的主要贡献有 2 方面:

1) 研究了基于时序分割网络的双流卷积神经网络,从网络输入数据的角度展开研究,提出了融合稀疏特征的时序分割网络,更好地聚焦运动目标.

2) 针对特征利用率低的问题,从网络结构的角度展开研究,提出了多层特征融合的行为识别时序分割网络,更好地融合特征.

1 相关工作

1.1 双流卷积神经网络

双流卷积神经网络分为空间流卷积神经网络和时间流卷积神经网络,这 2 个卷积神经网络分别处理视频的空间维度和时间维度,分别提取空间信息和时间信息,双流卷积神经网络基本结构如图 1 所示.其中,空间信息是指视频中的场景、物体等信息;时间信息是指视频中物体的运动信息.

空间流卷积神经网络的输入是单帧的 RGB 图像,能有效地识别静止图像中的人体行为动作,网络结构类似于常用的图像分类网络,通常使用 Alexnet,VGG16,GoogleNet 等深度模型作为空间流卷积神经网络,一般先在 ImageNet 上预训练,然后再将预训练的参数迁移到空间流网络中来,以提升网络训练速度和性能.时间流卷积神经网络的输入是堆叠的连续帧光流图像,光流能够表示视频中物体的运动信息,是利用像素点在时域上的变化以及其在

连续帧上的相关性来表示物体运动的一种方式.利用光流的这一特性,能有效识别连续帧之间的人体行为动作,为了时空网络融合时特征维度相匹配,时间流网络结构通常和空间流卷积网络相同.

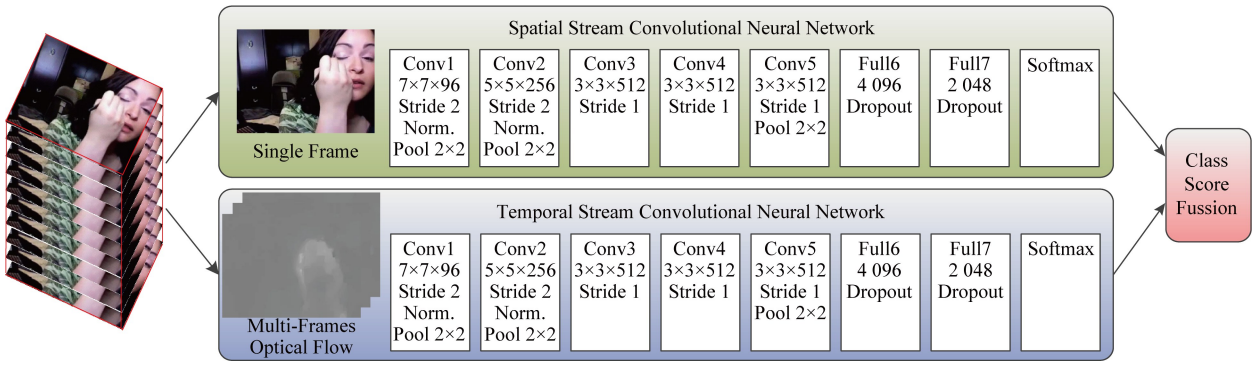


Fig. 1 Two-stream convolutional neural network

图 1 双流卷积神经网络结构

双流网络的融合是指空间流网络与时间流网络之间的融合,一般分为 2 种形式:1)空间流和时间流 2 个独立卷积网络在它们的 Softmax 层后进行融合,只是结果的融合,通常使用平均法和加权法融合不同类别的得分,得到最后的结果;2)时空网络在中间特征层融合.一般在某一网络层进行时空特征融合后形成混合的时空卷积网络;另一种融合方式是形成混合的时空卷积网络之后,保留单纯的空间流卷积网络或者时间流卷积网络,Softmax 层之后再一次融合不同类别的得分,得到最终的结果.

1.2 3D 卷积神经网络

在视频序列中使用卷积神经网络,最直接的方法是对视频序列的每一帧图像使用卷积神经网络来识别,但是这样对单帧图像的处理没有考虑连续帧之间的信息,在行为识别中行为的发生一般会持续一个过程,在连续帧之间存在运动信息.那么,为了有效利用连续帧之间的运动信息,文献[15]提出一种 3D 卷积神经网络的方法,即在卷积神经网络结构中采用 3D 卷积核进行卷积,3D 卷积核与 2D 卷积核相比,增加了时间维度,可以同时获取时间和空间维度的特征,在行为识别特征表示方面优于 2D 卷积.2D 卷积是在单帧图像的基础上进行卷积,通常选取 3×3 大小的卷积核,将 2D 卷积应用于图像将输出图像,将 2D 卷积应用于多个图像,将它们视为不同的通道,也会得到图像.因此,2D 卷积网络在每次卷积操作之后都会丢失输入信号的时间信息.3D 卷积是在相邻的几帧图像上进行卷积,卷积核大小一般为 3×3×3,只有 3D 卷积才能保留输入信号的时间信息,如图 2 所示:

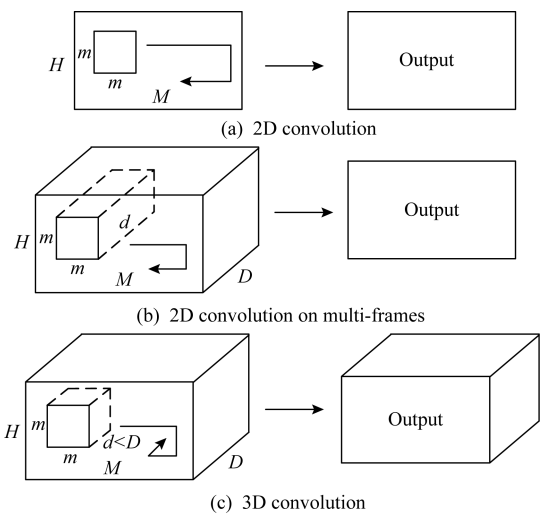


Fig. 2 2D convolution and 3D convolution

图 2 2D 卷积与 3D 卷积

3D 卷积神经网络体现时间维度是将多个连续的图像帧堆叠在一起,形成 1 个立方体,然后在立方体中使用 3D 卷积核进行卷积,卷积核的深度要小于堆叠的图像帧的数量.因此,3D 卷积中的每一个特征都会有相邻帧的特征相连,在连续帧上的表示便能获取到视频中物体的运动信息.

1.3 时序分割网络

给定 1 个视频 V,把它分成 K 段 $\{S_1,S_2,\cdots,S_K\}$,每段的时长相等.那么时序分割网络可以表示为

$$Q_{\mathrm{TSN}}(T_1,T_2,\cdots,T_K)=H(g(F(T_1;\boldsymbol{W}),F(T_2;\boldsymbol{W}),\cdots,F(T_K;\boldsymbol{W}))), \tag{1}$$

其中,$(T_1,T_2,\cdots,T_K)$ 是视频 V 中的单一帧组成的序列,而 T_k 是由其对应的视频子片段 S_k 中的帧随机采样产生,$k\in\{1,2,\cdots,K\}$;$F(T_k;\boldsymbol{W})$ 是输入

属于不同类别的分数预测函数,即视频帧 T_k 经参数为 $\boldsymbol{W}$ 的卷积神经网络后得到 1 个 C 维的向量,其表示 T_k 分别属于 C 类行为动作的预测得数;$g(\cdot)$ 是段共识函数,将多个子视频经卷积神经网络得到的预测结果进行融合,获得关于视频所属类别的一致性预测结果 $\boldsymbol{G}=(G_1,G_2,\cdots,G_C)^{\mathrm{T}}$,$C$ 表示类别数;基于以上的一致性预测结果,使用函数 $H(\cdot)$ 预测整个视频属于每个行为类别的概率,这里 $H(\cdot)$ 使用 Softmax 函数,概率最高的类别就是视频 V 所属类别.结合分类常用的交叉熵损失,最终视频 V 的类别预测损失函数可以表示为

$$L(y,\boldsymbol{G})=-\sum_{i=1}^{C}y_i(G_i-\sum_{j=1}^{C}\exp G_j),\quad(2)$$

其中,y_i 表示类别 i 的真值.这种时序分割网络是可微的,或者至少有次梯度的,由函数 $g(\cdot)$ 的选择决定,可以用反向传播算法和多个子视频帧来联合优化模型参数 $\boldsymbol{W}$.在反向传播过程中,模型参数 $\boldsymbol{W}$ 关于损失值 L 的梯度为

$$\frac{\partial L(y,\boldsymbol{G})}{\partial \boldsymbol{W}}=\frac{\partial L}{\partial \boldsymbol{G}}\sum_{k=1}^{K}\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{F}(T_k)}\frac{\partial \boldsymbol{F}(T_k)}{\partial \boldsymbol{W}},\quad(3)$$

其中,K 是 TSN 使用的子视频段数.TSN 从整个视频中学习模型参数而不是 1 个短的片段.与此同时,通过对所有视频固定 K,采用了一种稀疏时间采样策略,其中采样片段只包含一小部分帧.与先前使用密集采样帧的方法相比,这种方法大大降低计算开销.时序分割网络结构如图 3 所示:

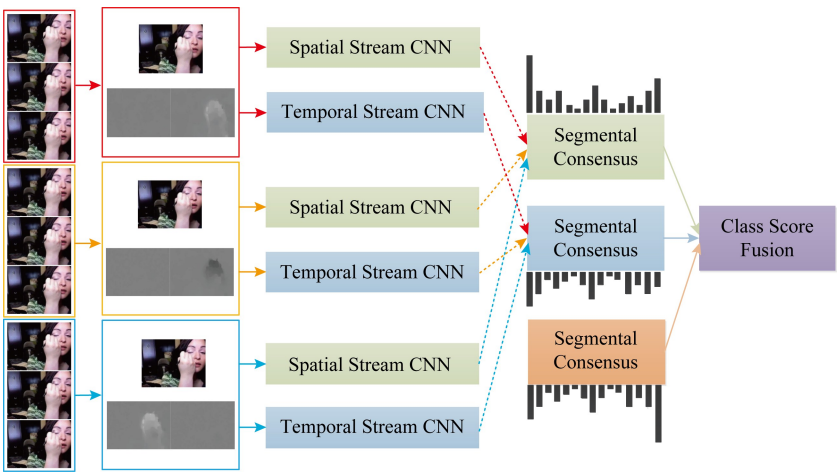


Fig. 3 Temporal segment network

图 3 时序分割网络结构

2 特征融合时序分割网络的行为识别

本节将详细从网络输入数据和网络结构 2 个方面展开研究:1)研究了融合稀疏特征的网络输入数据,目的是稀疏特征聚焦于视频中的前景目标,能够有效地提取图像中的运动物体,减少冗余信息,与 RGB 图像和光流图像包含的信息形成互补;2)利用卷积神经网络可视化验证了浅层卷积能提取细节特征,深层卷积能提取语义特征,将深度网络中高层特征的语义信息和低层特征的细节信息相结合,利用不同卷积层之间的特征优势互补,有利于网络捕捉人体行为的整体特征和不同类别之间的细节特征,从而提升行为识别的准确率.图 4 为算法的流程图.具体步骤为:1)将输入的视频平均分为 3 个子视频,对 3 个子视频随机采样,获取样本的 RGB、光流以及稀疏图像,分别输入到卷积网络中;2)提取各数据类型不同卷积层的特征,将卷积网络提取的特征按照不同的样本类型进行融合;3)利用 Softmax 函数进行行为分类.

2.1 稀疏特征

许多实际应用中已知的数据矩阵 $\boldsymbol{D}$ 往往是低秩或近似低秩的,但存在随机幅值任意大且分布稀疏的误差破坏了原有数据的低秩性,为了恢复矩阵 $\boldsymbol{D}$ 的低秩结构,可将矩阵 $\boldsymbol{D}$ 分解为 2 个矩阵之和,即 $\boldsymbol{D}=\boldsymbol{A}+\boldsymbol{E}$,其中矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{E}$ 未知,但 $\boldsymbol{A}$ 是低秩的,$\boldsymbol{E}$ 是稀疏的.

当矩阵 $\boldsymbol{E}$ 的元素服从独立同分布的高斯分布时,可用经典的主成分分析方法来获得最优的矩阵 $\boldsymbol{A}$,即转换为最优化问题:

$$\min_{\boldsymbol{A},\boldsymbol{E}}\|\boldsymbol{E}\|_{\mathrm{F}},\ \text{s.t. rank}(\boldsymbol{A})\leqslant r,\boldsymbol{D}=\boldsymbol{A}+\boldsymbol{E},\quad(4)$$

其中,$\|\cdot\|_{\mathrm{F}}$ 表示矩阵的 Frobenius 范数.

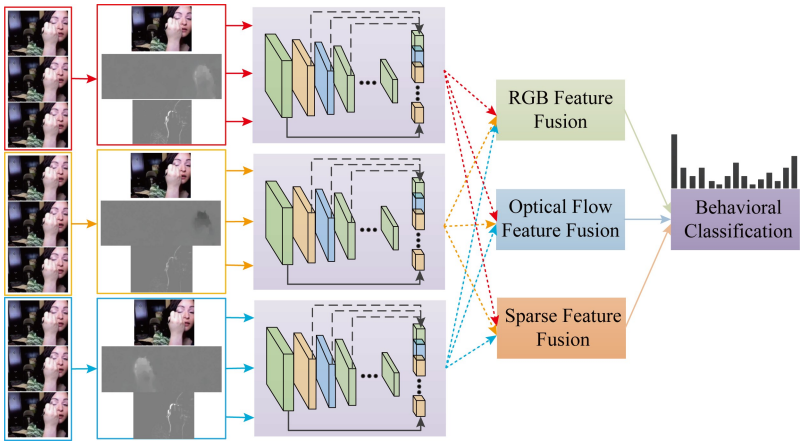


Fig. 4 Diagram of action recognition of temporal segment network based on feature fusion

图 4 特征融合时序分割网络的行为识别框图

当 $\boldsymbol{E}$ 为稀疏的大噪声矩阵时,PCA 无法给出理想的结果,可用鲁棒性主成分分析(robust principal component analysis, RPCA)来获取最优矩阵 $\boldsymbol{A}$,则式(4)问题可以转化为优化问题:

$$\min_{\boldsymbol{A},\boldsymbol{E}} \operatorname{rank}(\boldsymbol{A})+\lambda\|\boldsymbol{E}\|_0,\ \text{s.t.}\ \boldsymbol{D}=\boldsymbol{A}+\boldsymbol{E}, \tag{5}$$

其中秩函数 rank(•)、矩阵的 0 范数均非凸,变成了 NP-hard 问题,需要对其松弛.因为,核范数是秩函数的凸包,且 1 范数是 0 范数的凸包,故式(5)的 NP-hard 问题松弛后可转化为凸优化问题:

$$\min_{\boldsymbol{A},\boldsymbol{E}} \|\boldsymbol{A}\|+\lambda\|\boldsymbol{E}\|_1,\ \text{s.t.}\ \boldsymbol{D}=\boldsymbol{A}+\boldsymbol{E}, \tag{6}$$

其中,$\boldsymbol{A}$ 是低秩分量,$\boldsymbol{E}$ 是与其对应的稀疏分量;$\|\cdot\|_*$表示矩阵的核范数,是矩阵奇异值的和,同时也是 rank(•)的凸近似;$\|\cdot\|_1$表示 $L1$ 范数,λ 是 1 个大于零的加权参数,用来平衡 2 个范数.在一定条件下已经证明,只要误差矩阵 $\boldsymbol{E}$ 相对于矩阵 $\boldsymbol{A}$ 足够稀疏,就可以通过求解凸优化问题(式(4)),准确地从矩阵 $\boldsymbol{D}$ 中恢复低秩分量和稀疏分量,即最小化上述核范数和 $L1$ 范数的加权组合.

对于式(6)所描述的 RPCA 问题,可以使用增广拉格朗日乘子法来优化,拉格朗日函数为

$$L(\boldsymbol{A},\boldsymbol{E},Y,\mu)=\|\boldsymbol{A}\|_*+\lambda\|\boldsymbol{E}\|_1+\langle Y,\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}\rangle+\frac{\mu}{2}\|\boldsymbol{D}-\boldsymbol{A}-\boldsymbol{E}\|_{\mathrm{F}}^2, \tag{7}$$

其中,Y 为拉格朗日乘子,μ 是 1 个较小的正数.

RPCA 在图像和视频处理方向应用广泛,常用于图像矫正、去噪、视频背景建模与前景目标提取等方面,类似地,还有图像分割、显著性检测等[38-42].对于视频中的前景目标分割,由于帧与帧之间的相关性,背景被近似为低秩分量;而前景目标只占据图像中一小部分像素,例如人体运动,运动的人体部分可以看作是稀疏分量.通过以上的增广拉格朗日乘子法求解 RPCA 问题,对于行为动作视频可以得到如图 5 所示的稀疏特征.图 5 中第 1 行表示 RGB 图像,第 2 行表示 x 轴方向的行为运动光流图像,第 3 行表示 y 轴方向的行为运动光流图像,第 4 行表示稀疏图像.由图 5 可知,RGB 图像表示图像的表观

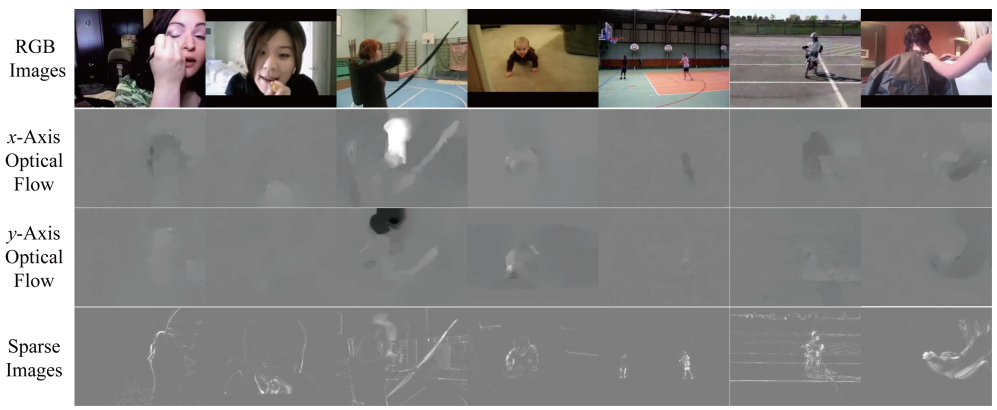


Fig. 5 Comparison of RGB, optical flow and low rank data

图 5 RGB、光流和低秩数据对比

特征,既包括背景,也包括前景目标;光流图像表示图像中运动物体的运动方向与速度,对于 x 轴方向,白色表示往右边运动,灰度值越高说明运动速度越快,黑色表示往左边运动,灰度值越低说明运动速度越快,其余灰色区域表示没有运动物体,y 轴方向同理,白色表示往上边运动,黑色表示往下运动;而稀疏特征图像不同于彩色和光流图像,其聚焦前景目标的行为动作,能有效地提取出运动物体,同时去除背景能有效降低数据的冗余度,显著提升网络训练速度.

2.2 网络特征融合

针对其网络中缺乏信息交互的缺点,将深度网络中高层语义信息和低层细节信息结合起来共同识别行为动作,使网络性能更具优势.多层特征融合是建立在卷积神经网络低层细节特征和高层语义特征基础之上的,利用不同深度卷积层特征具备的特点来实现.以 inceptionv2 网络为例来说明改进后的卷积神经网络工作原理,如图 6 所示.该网络是由多流卷积神经网络组合而成.对于空间流卷积神经网络而言,假设输入的彩色图像尺寸大小为 224×224×3,首先选取尺寸大小为 7×7、步长为 2 的卷积核,利用卷积层提取输入图像的特征,得到 64 个大小为 112×112 的特征图,然后进行最大池化得到 56×56 的特征图;选取尺寸大小为 3×3、步长为 2 的卷积核,再次卷积提取池化后的特征并二次池化,得到池化后的特征大小为 28×28×192.接着,将得到的特征依次经过 10 个 inception 结构单元,分别是从结构单元 inception3a 到 inception5b,得到的特征大小为 7×7×1024,再次经过 1 个平均池化输出 1×1×1024 的特征,展开为 1D 向量作为全连接层的输入之一;与此同时,将浅层卷积后的输出特征也展开为 1D 向量送入全连接层.最后,以浅层卷积特征和深层卷积特征一同输入全连接层,形成 1×101 的向量.

如图 6 所示,以融合 inception3a 层的输出特征和 inception5b 的输出特征为例来说明多层卷积特征融合过程.为了清楚说明高低维度特征的融合原理,表 1 列出了卷积神经网络各层输出的特征尺寸大小.

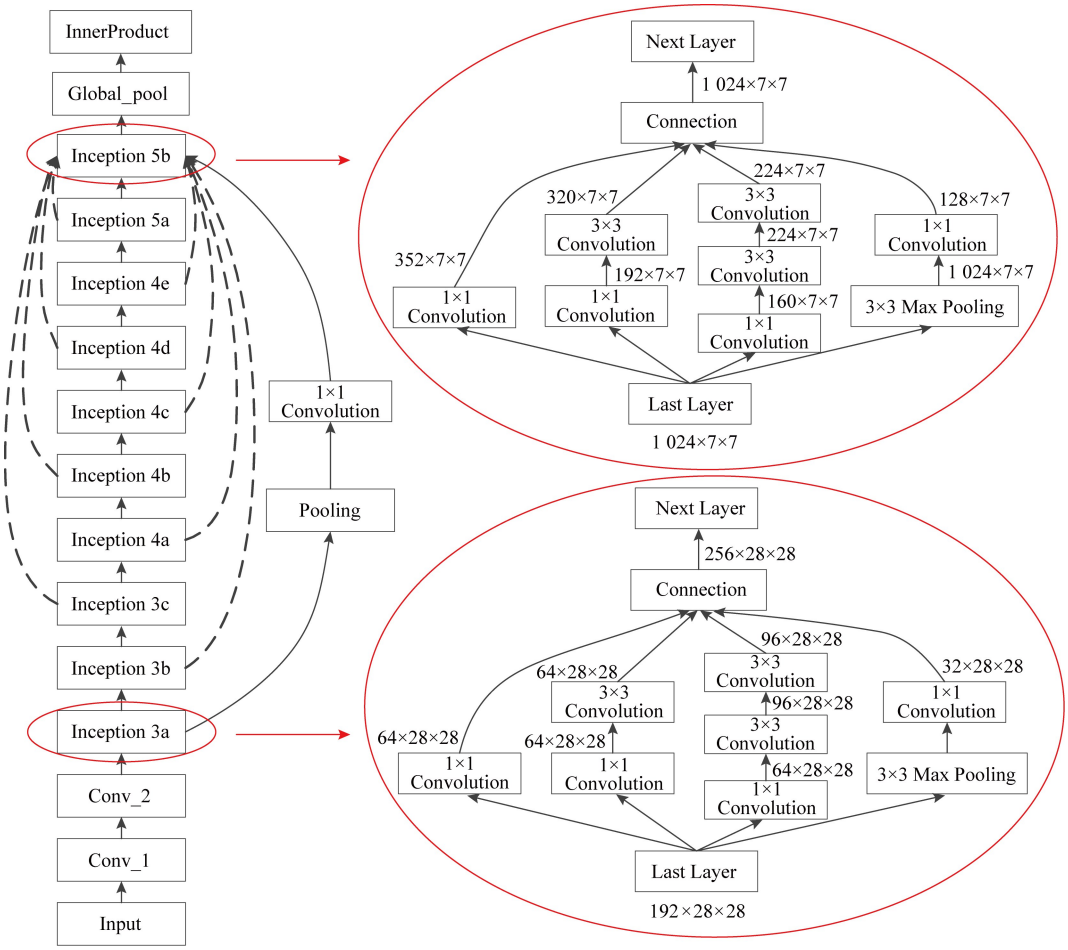


Fig. 6 Diagram of multi-layer convolution feature neural network

图 6 多层卷积特征神经网络示意图

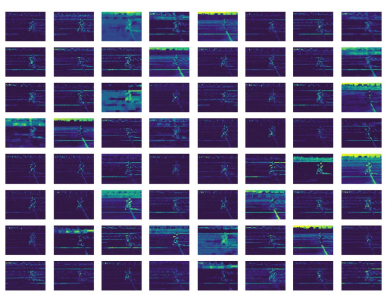
Table 1 Map Size of Network Layers

表 1 网络各层输出特征图尺寸

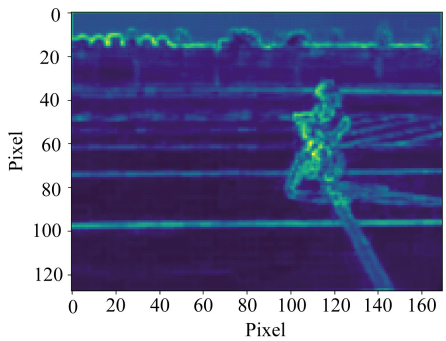
Network Layers	Kernel Size/Stride	Output Size
Convolution_1	7×7/2	112×112×64
Pooling	3×3/2	56×56×64
Convolution_2	3×3/1	56×56×192
Pooling	3×3/2	28×28×192
Inception3a		28×28×256
Inception3b		28×28×320
Inception3c	2	28×28×576
Inception4a		14×14×576
Inception4b		14×14×576
Inception4c		14×14×576
Inception4d		14×14×576
Inception4e	2	14×14×1 024
Inception5a		7×7×1 024
Inception5b		7×7×1 024
Pooling		1×1×1 024

首先,输入图像经过前 2 层卷积层和池化层之后得到 28×28×192 的特征图,前 2 维数据表示特征图的长和宽,第 3 维数据表示通道数.然后,将特征送入 inception3a 层,经过 inception 结构单元中的 4 个支路分别得到 4 组特征,将这 4 组特征串联起来作为下一层的输入.与此同时,对该特征进行池化操作,这里选择平均池化,相较与最大池化,平均池化在减少维度的同时,能够保留更多的图片背景信息,有利于信息传递到下一个模块进行特征提取,并且使得其尺寸与深层卷积特征尺寸相同,便于特征融合.另外,由于特征融合会增加特征维度,增大计算复杂度,通过卷积核为 1×1 的卷积做降维,得到浅层卷积特征.将浅层卷积特征与 inception5b 层的输出特征串联起来,展开为 1 维向量作为全连接层的输入.

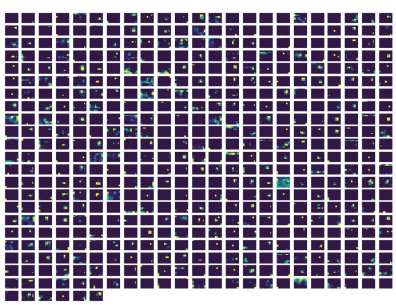
时间流卷积网络和稀疏卷积神经网络与空间流卷积神经网络类似,按照上述网络得到浅层卷积特征,与最后一层 inception 结构单元输出的深层卷积特征融合参与最终的分类工作.对于 2 个特征映射 $x_t^a \in \mathbb{R}^{H\times M\times D}$ 和 $x_t^b \in \mathbb{R}^{H'\times M'\times D'}$,要使用它们生成特征映射 $y_t \in \mathbb{R}^{H''\times M''\times D''}$ 方式多样,其中 t 表示时间,H,M,D 分别表示 3 个特征图各自的高度、宽度和通道数量.由于串联融合简单高效,本文采用串联融合的方式将低层细节信息和高层语义信息进行融合,低层细节信息主要提取的是颜色、纹理等细节特征,如图 7(a)(b)所示;而高层语义信息更具有代表性,越往后越能确定网络提取了图中哪部分的特征用来进行行为识别,如图 7(c)(d)所示,通过融合可以充分利用特征,形成信息互补,提升识别准确率.



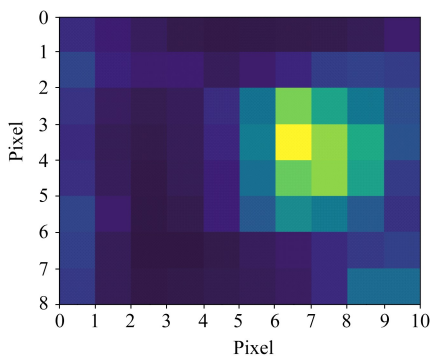
(a) Visualization of low-level convolution feature map



(b) Random image of (a)



(c) Visualization of high-level convolution feature map



(d) Random image of (c)

Fig. 7 Feature visualization of convolutional neural networks

图 7 卷积神经网络特征可视化

串联融合即将 2 个特征对应通道的特征图按照先后顺序串联在一起,以组合后的特征作为新的特征,即:

$$\begin{cases} y_{i,j,2d} = x^{a}_{i,j,d}, \\ y_{i,j,2d-1} = x^{b}_{i,j,d}, \end{cases} \tag{8}$$

其中,$1 \leqslant i \leqslant H$,$1 \leqslant j \leqslant M$,$1 \leqslant d \leqslant D$,并且 $x^{a}, x^{b} \in \mathbb{R}^{H \times M \times D}$,$y \in \mathbb{R}^{H \times M \times D}$.

3 实验与仿真

本节在 2 个大型动作数据集上进行实验,来验证融合特征的时序分割网络的有效性.这 2 个数据集分别是 UCF101 和 HMDB51.其中,UCF101 数据集包含 101 个动作类别和 13 320 个视频剪辑;HMDB51 数据集是包括各种来源(如电影和网络视频)的大量逼真视频,数据集由来自 51 个动作类别的 6 849 个视频剪辑组成.实验遵循数据集原始评估方案,使用 3 个训练/测试分组,分别为数据集分组 1、数据集分组 2、数据集分组 3,并以这些分组的平均准确率作为最终的行为识别准确率.

本节实验使用小批量随机梯度下降算法来学习网络参数,批量大小设置为 32,动量设置为 0.9.另外,使用数据集预先训练的模型初始化网络权重,并在实验中设置了较小的学习率.对于空间网络,学习速率初始化为 0.001,并且每 2 000 次迭代减少 1/10,整个训练过程在 10 000 次迭代时停止;对于时间网络和稀疏网络,将学习速率初始化为 0.005,在 12 000 和 18 000 次迭代后减少到 1/10,最大迭代次数设置为 20 000.为了快速提取光流,选择在 OpenCV 中用 CUDA 实现的 TVL1 光流算法;为了加速训练,采用了多个 GPU 的数据并行策略,使用 Caffe 和 OpenMPI2 实现.

3.1 实验数据集

UCF101 有 13 320 个视频,包括 101 个动作类别,动作多样性大,在摄像机运动、物体外观和姿态、物体比例、视点、杂乱背景、光照条件等方面存在较大变化,并且动作视频均由剪辑而成,而非演员表演,具有一定的真实性,是目前最具挑战性的数据集,数据集中部分动作类别如图 8 所示:

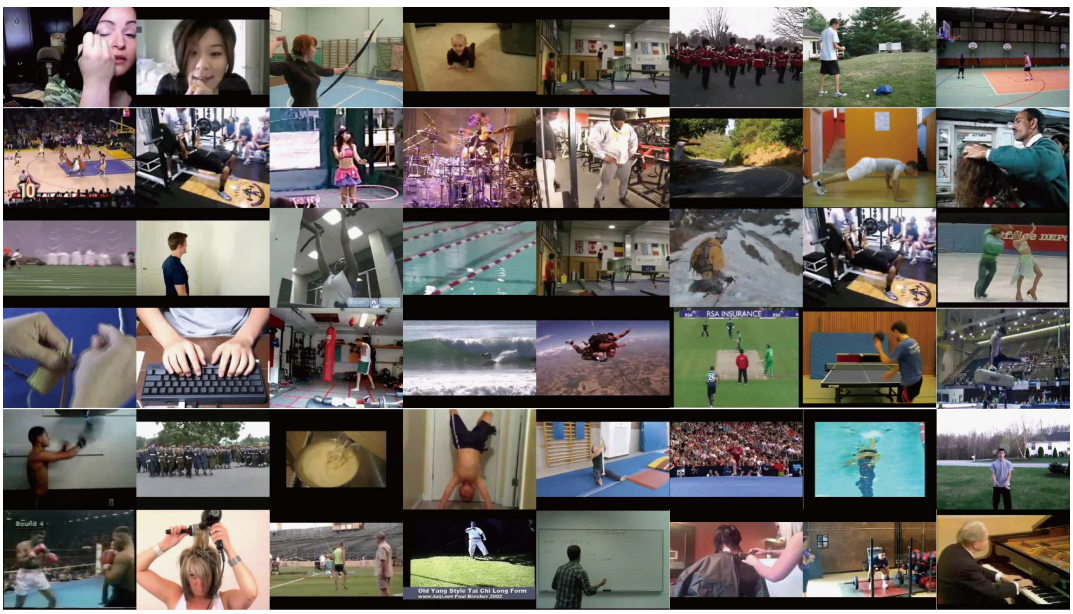


Fig. 8 Partial action categories in UCF101 dataset

图 8 UCF101 数据集部分动作类别

HMDB51 数据集包含 6 766 个视频剪辑,分为 51 个动作类别,每个动作类别至少包含 101 个视频剪辑,部分动作类别如图 9 所示.HMDB51 数据集大部分来自电影,一小部分来自公共视频网站,如 Prelinger archive,YouTube 和谷歌视频等.

3.2 实验环境

深度学习硬件环境:CPU E5-2696v4,GPU 2 块 GTX1080Ti,256 GB 固态硬盘,32 GB 内存;网络的学习与测试环境:Ubuntu16.04,NVIDIA CUDA 8.0,cudnnv5,Caffe,opencv3.0,Python.

3.3 稀疏特征引入效果

实验在 UCF101 和 HMDB51 这 2 个公共行为识别数据集上对其进行了验证,并与近年来一些经典算法以及常用算法进行了比较,对比结果如表 2 所示:

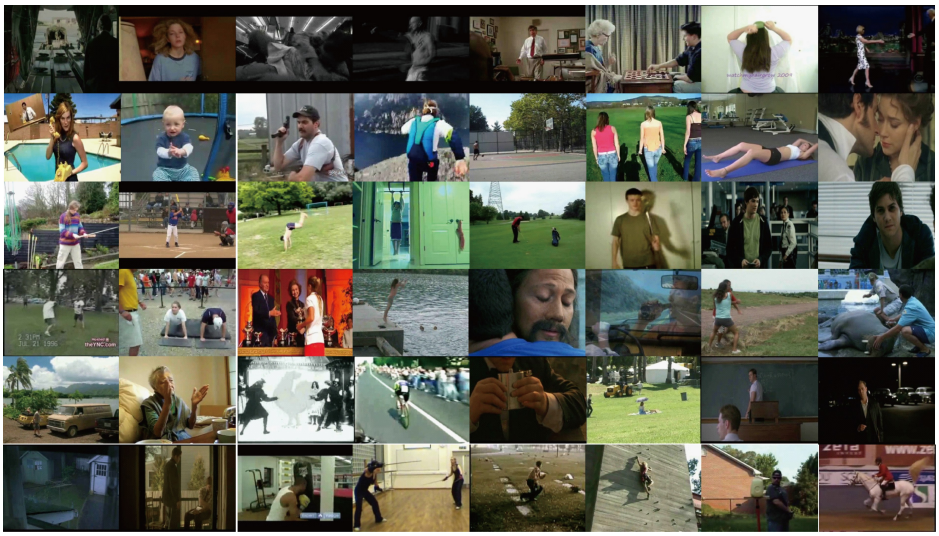


Fig. 9 Partial action categories in HMDB51 dataset

图 9 HMDB51 数据集部分动作类别

Table 2 Accuracy Comparison of Different Algorithms on UCF101 and HMDB51 Datasets

表 2 UCF101 和 HMDB51 数据集上不同算法准确率比较 %

Algorithms	UCF101	HMDB51
DT+MVSV[43]	83.5	55.9
IDT+FV[44]	85.9	57.2
IDT+HSV[45]	87.9	61.1
MoFAP[46]	88.3	61.7
C3D+IDT[15]	90.4	
TDD+IDT[47]	91.5	65.9
LTC[48]	91.7	64.8
LTC+IDT[48]	92.7	67.2
P3D ResNet+IDT[33]	93.7	
Two Stream[20]	88.0	59.4
Two Stream+LSTM[49]	88.6	
Two Stream Fusion[16]	92.5	65.4
Transformations[50]	92.4	62.0
TSN(RGB+Optical Flow)[28]	94.0	69.2
Sparse+TSN	**96.9**	**76.4**

Note: Bold fonts represent the best values in different algorithms.

从表 2 可以看出,算法分为 3 类:

第 1 类是不使用深度学习算法的传统经典机器学习算法,该类算法手动提取行为特征,稳定性高,在 UCF101 数据集上识别率可达到 88%左右,在 HMDB51 数据集上识别率超过了 61%.例如,文献[46]中提出的一种名为 MoFAP 的组合运动特征,该特征由 3 部分组成:局部运动特征、运动原子、运动语句.其中,运动原子指运动过程中的某一子阶段,而运动语句就是这些子阶段的组合,例如跳高分为 3 个子阶段,助跑、起跳和着陆,即运动原子;三者之间的不同组合就成为运动语句,通过这种方式使得特征对行为的表征能力更强,以提高识别精度.

第 2 类是使用 3D 卷积的深度学习算法,该类算法速度快,可以达到实时,且识别率较传统算法高出 4%以上.例如,文献[48]认为不同的动作具有不同的时间和空间模式,有些行为可能需要长时间的行为动态才能辨认,提出了 LTC 网络结构,通过增加输入视频的时长以提高识别准确率.

第 3 类是使用双流卷积神经网络的算法,该类算法精度最高,可以达到 94%以上.由表 2 可知,融合稀疏特征的时序分割网络较时序分割网络有一定提升,在 UCF101 上识别率可达到 96%以上,在 HMDB51 上识别率超过了 76%.

3.4 多层特征融合实验

为了验证多层卷积特征融合卷积网络的有效性,以 UCF101 数据集分组 1 的实验为例,从结构单元 inception3a 层到 inception5a 层的输出与 inception5b 层的特征进行融合,各层融合之后的网络识别率.表 3 列出了 RGB、光流图像和稀疏图像训练的时序分割网络在加入多层特征融合方法之后的识别率.与 RGB 图像类似,利用光流图像和稀疏图像训练的时序分割网络也是在 incepteion5a 层输出的特征和 incepteion5b 层输出的卷积特征融合后,得到的识别率最高,分别达到了 93.56%和 86.10%,光流基本维持不变,稀疏网络较不融合浅层卷积特征的网络识别率高了 0.6%以上,说明了浅层特征的加入对于网络性能的改善.

Table 3 Comparison Recognition Rate of Different Convolution Layers Fusion under UCF101 Dataset

表 3 UCF101 数据集分组 1 下不同卷积层融合的识别率对比

%

Fusion Layer	RGB	Optical Flow	Sparse
Inception3a→Inception5b	87.70	92.59	85.02
Inception3b→Inception5b	87.85	93.03	84.80
Inception3c→Inception5b	87.83	92.98	84.94
Inception4a→Inception5b	87.85	92.86	85.36
Inception4b→Inception5b	87.13	92.91	85.86
Inception4c→Inception5b	87.10	92.77	85.45
Inception4d→Inception5b	87.71	92.99	85.71
Inception4e→Inception5b	88.09	92.77	85.97
Inception5a→Inception5b	88.22	93.56	86.10

为了进一步验证多层特征融合的行为识别时序分割网络的有效性,实验在 UCF101 和 HMDB51 这 2 个公共行为识别数据集上对其进行了验证,并与近年来一些经典算法以及常用算法进行了比较,对比结果如表 4 所示.

从表 4 可以看出,多层特征融合的行为识别时序分割网络较原有的融合稀疏特征的时序分割网络有一定的提升,UCF101 识别率为 97.1%,在 HMDB51 数据集上可以达到 76.7%,说明浅层卷积层与深层卷积融合对于网络性能的提升具有一定的作用.其准确率混淆矩阵图 10 和图 11 所示,x 轴表示预测的视频动作类别,y 轴表示真实的视频动作类别,右

Table 4 Accuracy Comparison of Different Algorithms on UCF101 and HMDB51 Datasets

表 4 UCF101 和 HMDB51 数据集上不同算法准确率比较

%

Algorithms	UCF101	HMDB51
C3D+IDT[15]	90.4	
TDD+IDT[47]	91.5	65.9
LTC[48]	91.7	64.8
LTC+IDT[48]	92.7	67.2
P3D ResNet+IDT[33]	93.7	
Two Stream[20]	88.0	59.4
Two Stream+LSTM[49]	88.6	
Two Stream Fusion[16]	92.5	65.4
Transformations[34]	92.4	62.0
TSN(RGB+Optical Flow)[28]	94.0	69.2
Multi-layer Feature Fusion	**97.1**	**76.7**

Note: Bold fonts represent the experimental results of our method.

侧图例颜色越深表示准确率或者误识率越高,颜色越浅表示准确率或者误识率越低;混淆矩阵对角线上的小方块表示识别准确率,其余位置的小方块表示误识率,即视频实际属于小方块所在行对应的类别,被误识为小方块所在列对应的类别;且每一行小方块对应的概率之和为 1,若该行对角线上的小方块对应的概率为 1,该类别识别准确率为 100%,若该行对角线上的小方块对应的概率小于 1,则该类别存在误识.例如 UCF101 数据集中,第 80 和第 81 个类别

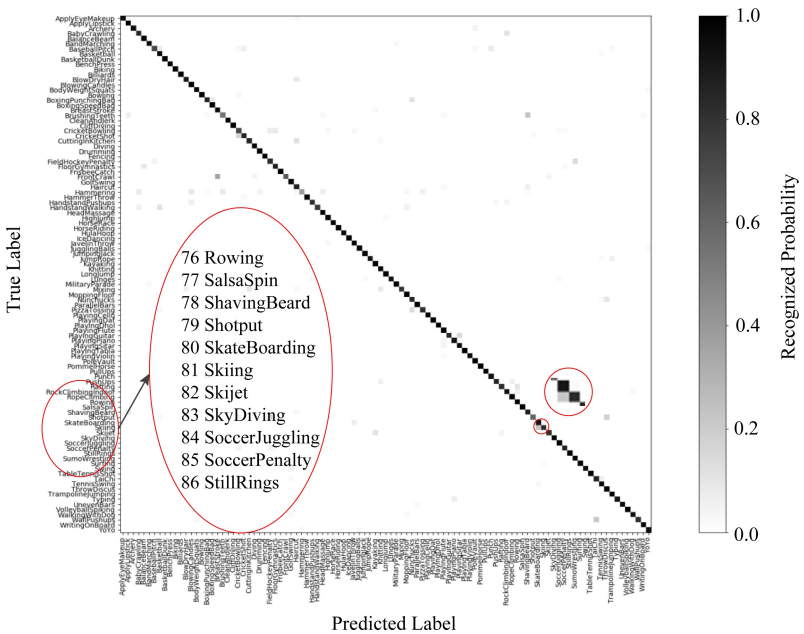


Fig. 10 Accuracy confusion matrix on UCF101 dataset

图 10 UCF101 数据集准确率混淆矩阵

分别为滑板和滑雪,如图 12 所示,分别例举了其 RGB 图像、光流图像和稀疏图像,可以看出这 2 个动作类别较为相似,观察混淆矩阵中局部放大部分可以发现,这 2 个类别的误识率相对于其他类别偏高.

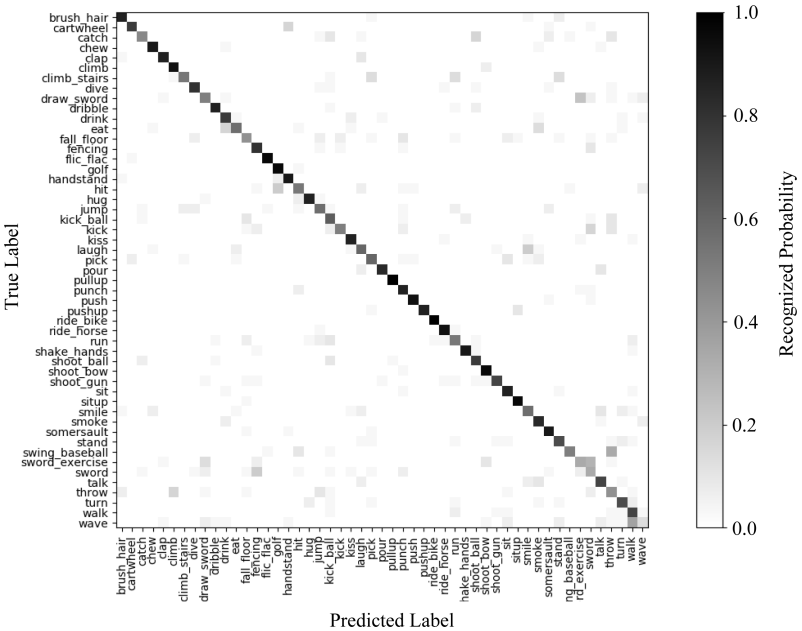


Fig. 11 Accuracy confusion matrix on HMDB51 dataset

图 11 HMDB51 数据集准确率混淆矩阵

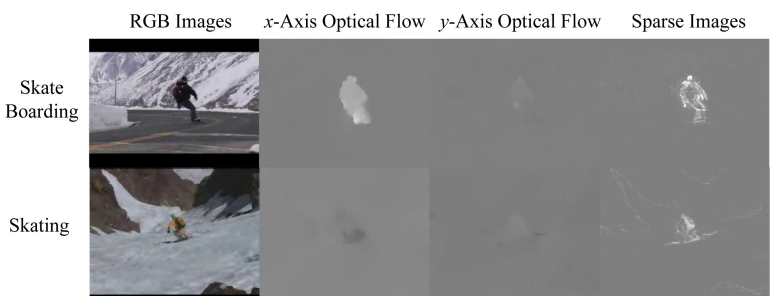


Fig. 12 Legend of the 80th and 81st categories

图 12 第 80 个和第 81 个类别图例

4 总 结

本文研究了基于时序分割网络的双流卷积神经网络,并在它的基础上提出了融合稀疏特征的时序分割网络.同时针对特征利用率低的问题,提出了多层特征融合的行为识别时序分割网络.本文基于稀疏特征和多层特征融合的行为识别网络,算法在公共库 UCF101 和 HMDB51 上的识别效果要好于主流算法.

人体动作识别是一项具有挑战的任务,本文提出特征融合时序分割网络的行为识别,从网络输入数据的角度展开研究,虽然在一定程度上和 RGB 图像以及光流图像表示的特征存在互补,但单独使用时,效果均不如 RGB 图像和光流图像,如何优化稀疏特征,减少矩阵分解过程中的信息缺失,提高它的表征能力,是需要进一步研究.针对特征利用率低的问题,从网络结构的角度展开研究,提出了多层特征融合的行为识别时序分割网络,主要研究了浅层特征与深层特征的融合,虽然一定程度上提高了特征利用率,但是这还远远不够,不同网络之间的特征交互是需要进一步研究.目前,大多数行为识别方法都使用光流来表示运动特征,但光流提取耗时耗力,并且光流所包含的运动特征未必是最优特征,寻找优质的运动特征替代光流,提升行为识别效率,这些需要进一步研究和探索.

参 考 文 献

[1] Yao Guangle, Lei Tao, Zhong Jiandan. A review of convolutional-neural-network-based action recognition [J]. Pattern Recognition Letters, 2019, 118(1): 14-22

[2] Shan Yanhu, Zhang Zhang, Huang Kaiqi. Visual human action recognition: History, status and prospects [J]. Journal of Computer Research and Development, 2016, 53(1): 93-112 (in Chinese)
(单言虎, 张彰, 黄凯奇. 人的视觉行为识别研究回顾、现状及展望[J]. 计算机研究与发展, 2016, 53(1): 93-112)

[3] Lei Chen, Song Zhanjie, Lu Jiwen, et al. Learning principal orientations and residual descriptor for action recognition [J]. Pattern Recognition, 2019, 86(2): 14-26

[4] Hao Yazhou, Zheng Qinghua, Chen Yanping, et al. Recognition of abnoamal behavior based on data of public opinion on the Web [J]. Journal of Computer Research and Development, 2016, 53(3): 611-620 (in Chinese)
(郝亚洲, 郑庆华, 陈艳平, 等. 面向网络舆情数据的异常行为识别[J]. 计算机研究与发展, 2016, 53(3): 611-620)

[5] Laptev I. On space-time interest points [J]. International Journal of Computer Vision, 2005, 64(2/3): 107-123

[6] Harris C J. A combined corner and edge detector [C] //Proc of the 4th Alvey Vision Conf. Berlin: Springer, 1988: 147-151

[7] Oikonomopoulos A, Patras I, Pantic M. Spatiotemporal salient points for visual recognition of human actions [J]. IEEE Transactions on Cybernetics, 2006, 36(3): 710-719

[8] Dollar P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features [C] //Proc of IEEE Int Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Piscataway, NJ: IEEE, 2006: 65-72

[9] Tong Ming, Wang Fan, Wang Shuo, et al. A new framework of action recognition: 3DHOGTCC and 3DHOOFG [J]. Journal of Computer Research and Development, 2015, 52(12): 2802-2812 (in Chinese)
(同鸣, 王凡, 王硕, 等. 一种3DHOGTCC和3DHOOFG的行为识别新框架[J]. 计算机研究与发展, 2015, 52(12): 2802-2812)

[10] Willems G, Tuytelaars T, Gool L V. An efficient dense and scale-invariant spatio-temporal interest point detector [C] // Proc of European Conf on Computer Vision. Berlin: Springer, 2008: 650-663

[11] Wang Heng, Alexander K, Schmid C, et al. Action recognition by dense trajectories [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2011: 3169-3176

[12] Murthy O V R, Goecke R. Ordered trajectories for human action recognition with large number of classes [J]. Image and Vision Computing, 2015, 42(10): 22-34

[13] Cho J, Lee M, Chang H J, et al. Robust action recognition using local motion and group sparsity [J]. Pattern Recognition, 2014, 47(5): 1813-1825

[14] Rahmani H, Mian A, Shah M. Learning a deep model for human action recognition from novel viewpoints [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(3): 667-681

[15] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks [C] //Proc of IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2014: 4489-4497

[16] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 1933-1941

[17] Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal multiplier networks for video action recognition [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7445-7454

[18] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(5): 436-444

[19] Ji Shuiwang, Xu Wei, Yang Ming, et al. 3D convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231

[20] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [J]. Neural Information Processing Systems, 2014, 1(4): 568-576

[21] Wang Heng, Schmid C. Action recognition with improved trajectories [C] // Proc of IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2014: 3551-3558

[22] Zhu Wangjiang, Hu Jie, Sun Gang, et al. A key volume mining deep framework for action recognition [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 1991-1999

[23] Kar A, Rai N, Sikka K, et al. AdaScan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 5699-5708

[24] Zhu Yi, Lan Zhenzhong, Newsam S, et al. Hidden two-stream convolutional networks for action recognition [J]. arXiv preprint arXiv:1704.00389, 2017

[25] Sevilla-Lara L, Liao Yiyi, Guney F, et al. On the integration of optical flow and action recognition [J]. arXiv preprint arXiv:1712.0416, 2017

[26] Zhang Bowen, Wang Limin, Wang Zhe, et al. Real-time action recognition with deeply-transferred motion vector CNNs [J]. IEEE Transactions on Image Processing, 2018, 27(5): 2326-2339

[27] Choutas V, Weinzaepfel P, Revaud J, et al. PoTion: Pose MoTion representation for action recognition [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7024-7033

[28] Wang Limin, Xiong Yuanjun, Wang Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2016: 20-36

[29] Lan Zhenzhong, Zhu Yi, Hauptmann A G, et al. Deep local video feature for action recognition [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2017: 1219-1225

[30] Zhou Bolei, Andonian A, Torralba A. Temporal relational reasoning in videos [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2018: 831-846

[31] Xu Huijuan, Das A, Saenko K. R-C3D: Region convolutional 3D network for temporal activity detection [C] //Proc of IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 5794-5803

[32] Ren Shaoqing, He Kaiming, 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

[33] Qiu Zhaofan, Yao Ting, Mei Tao. Learning spatio-temporal representation with pseudo-3D residual networks [C] //Proc of IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 5534-5542

[34] Gers F A, Schmidhuber J. Recurrent nets that time and count [C] // Proc of the IEEE-INNS-ENNS Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2000: 189-194

[35] Long Xiong, Gan Chuang, Gerard D M, et al. Multimodal keyless attention fusion for video classification [C] //Proc of Association for the Advancement of Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 7202-7209

[36] Du Wenbin, Wang Yali, Qiao Yu. RPAN: An end-to-end recurrent pose-attention network for action recognition in videos [C] //Proc of IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 3745-3754

[37] Wang Xiaolong, Girshick R, Gupta A, et al. Non-local neural networks [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7794-7803

[38] Kim H, Joonki P. Video summarization using low-rank sparse representation [J]. IEEE Transactions on Smart Processing and Computing, 2018, 7(3): 236-244

[39] Xie Wenbin, Yin Hong, Wang Meini, et al. Low-rank structured sparse representation and reduced dictionary learning-based abnormity detection [J]. IET Computer Vision, 2019, 13(1): 8-14

[40] Liu Xin, Zhao Guoying, Yao Jiawen, et al. Background subtraction based on low-rank and structured sparse decomposition [J]. IEEE Transactions on Image Processing, 2015, 24(8): 2502-2514

[41] Jin K H, Ye J C. Sparse and low-rank decomposition of a hankel structured matrix for impulse noise removal [J]. IEEE Transactions on Image Processing, 2018, 27(3): 1448-1461

[42] Zhang Xiujun, Xu Chen, Li Min, et al. Sparse and low-rank coupling image segmentation model via nonconvex regularization [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2015, 29(2): 1555004

[43] Cai Zhuowei, Wang Limin, Peng Xiaojiang, et al. Multi-view super vector for action recognition [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 596-603

[44] Wang Heng, Schmid C. LEAR-INRIA submission for the thumos workshop [C] //Proc of ICCV Workshop on THUMOS Challenge. Berlin: Springer, 2013: 1-3

[45] Peng Xiaojiang, Wang Limin, Wang Xingxing, et al. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice [J]. Computer Vision and Image Understanding, 2016, 150(9): 109-125

[46] Wang Limin, Qiao Yu, Tang Xiaoou. MoFAP: A multi-level representation for action recognition [J]. International Journal of Computer Vision, 2016, 119(3): 254-271

[47] Wang Limin, Qiao Yu, Tang Xiaoou. Action recognition with trajectory-pooled deep-convolutional descriptors [C] // Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 4305-4314

[48] Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1510-1517

[49] Ng Y H, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification [C] // Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 4694-4702

[50] Wang Xiaolong, Farhadi A, Gupta A. Actions transformations [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2658-2667

Li Hongjun, born in 1981. PhD, associate professor. Senior member of CAA and member of CCF. His main research interests include image processing, pattern recognition and artificial intelligence.

Ding Yupeng, born in 1993. Master. His main research interests include deep learning and image processing.

Li Chaobo, born in 1995. Master candidate. Her main research interests include computer vision and deep learning.

Zhang Shibing, born in 1962. PhD, professor, PhD supervisor. His main research interests include wireless communications, intelligent signal processing, machine learning, and cognitive radios.